

1971

A Variation of the Standard Steepest Ascent Search Procedure for Response Surface Experiments

George Whitfield Cobb
gcobb@mtholyoke.edu

Follow this and additional works at: <http://scholarscompass.vcu.edu/etd>

 Part of the [Biology Commons](#)

© The Author

Downloaded from

<http://scholarscompass.vcu.edu/etd/4443>

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

R111
M489
cobb
1971

A VARIATION OF THE STANDARD
STEEPEST ASCENT SEARCH PROCEDURE
FOR RESPONSE SURFACE EXPERIMENTS

by

George Whitfield Cobb

A.B., Dartmouth College, 1968

Thesis

submitted in partial fulfillment of the requirements for the
Degree of Master of Science in the Department of
Biometry at the Medical College of Virginia

Richmond, Virginia

June, 1971

4/

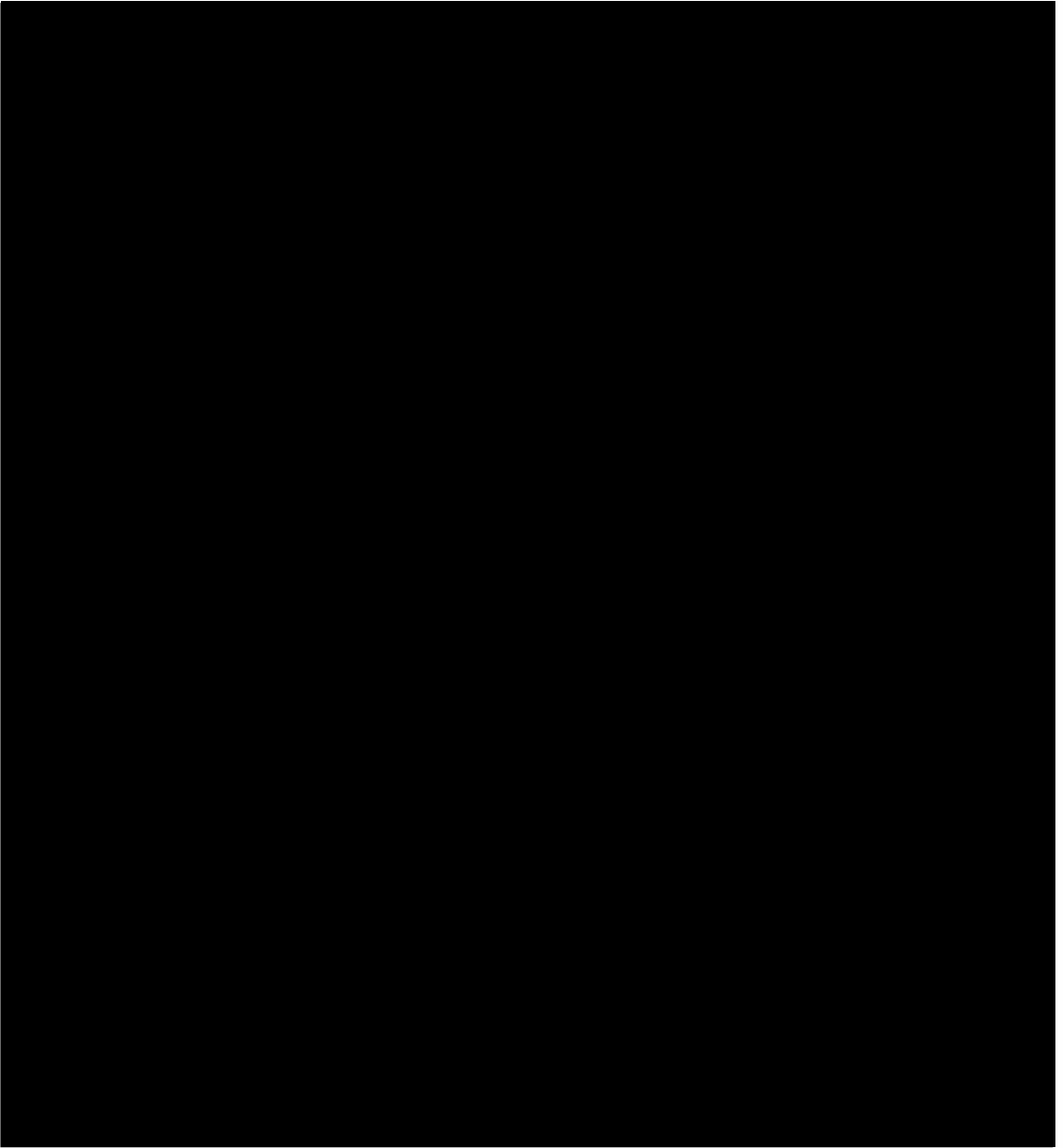
- 997

This thesis by George Whitfield Cobb is accepted in its present form as satisfying the thesis requirement for the degree of Master of Science.

Date:	Approved:
5/20/71.....	[Redacted].....
	Adviser, Chairman of Graduate Committee
5/20/71.....	[Redacted].....
5/20/71.....	[Redacted].....
5/20/71.....	[Redacted].....

APPROVE	[Redacted].....
	Dean of the School of Graduate Studies

CURRICULUM VITAE



PREFACE AND ACKNOWLEDGMENTS

This thesis examines the method of first-order steepest ascent as it is currently used in response surface experiments (Chapter I). Several possible improvements are considered (Chapter II), and one concrete proposal involving a new "stopping rule" is shown to lead to a significant increase in the efficiency of existing methods (Chapter III, IV).

I undertook this thesis with two separate goals, one institutional, one personal; each valid in its own right, yet neither wholly subordinate to the other. The institutional goal was to satisfy one of the requirements for the Master of Science degree, to demonstrate a greater or lesser competence in manipulating the archetypical facts of my chosen discipline by placing a small but in some sense acceptable contribution on the burgeoning heap of facts erected by my predecessors. The personal goal was to challenge the process we dignify with the name research, to observe first-hand the blend of frustration and exhilaration which one experiences in pursuing dead ends and uncovering new passageways, but which one inevitably buries beneath the discrete convention of third-person reporting. The institutional goal requires that this thesis be a concise recapitulation of those results relevant to the discipline, (though submitted in bureaucratically redundant quintuplicate); I

have attempted to meet this requirement. But the personal goal simultaneously requires that this thesis be more than an impersonal dessication of that which has held constant albeit capricious claim to my attention during the last six months; therefore I have included several paragraphs which chronicle the discarded directions and reveal the geneology of those results composing the thesis proper.

Special acknowledgment is due to the members of my graduate committee, Dr. Ray Myers (chairman), Dr. Roger Flora, and Dr. J. Van Bowen, for their assistance and efforts on my behalf; to the Department of Biometry, for the use of its IBM 1130 computer (supported by Grant No. 5 P07 RR00016-09 from the National Institute of Health), to Mr. John Howell for his most welcome assistance with the computer, and to Mrs. Frances Gokey and Mrs. Shirley Carpenter, who heroically rose to the challenge of typing a mathematical manuscript under immense pressure.

I particularly wish to thank my advisor, Dr. Ray Myers, for providing a truly outstanding source of sensitive guidance and encouragement. Without his guidance my efforts might have been diverted to peripheral problems; without his encouragement the dead ends encountered early in the project might have assumed more intimidating significance; and in either case the institutional goal would not have been realized. But without his sensitivity the mere completion of the thesis to satisfy an insitutional requirement would

have been a hollow accomplishment, devoid of the immense satisfaction which has accompanied the realization of my personal goal. Dr. Myers set for himself a difficult course, allowing me wherever possible to engineer my own successes and failures, while insuring, at the cost of frequent and lengthy conferences, that I never dwelt too long on the failures. In deepest appreciation of the experience he has given me, I dedicate this thesis to Ray Myers.

TABLE OF CONTENTS

	Page
I. A Context for Steepest Ascent	1
A. The Experimenter's Problem	1
B. The Statistician's Problem	6
C. Brief Description of Current Methods	8
D. Region Seeking Procedures	20
E. Qualitative Evaluation of First-Order Steepest Ascent	35
II. Early Dead Ends	41
A. Design for Wedge-Shaped Regions of Interest .	42
B. Weighted Least Squares	49
C. The Multivariate Integral Problem	68
III. The Stopping Rule Problem	70
A. Reparameterization and Problem Definition . .	71
B. The Current Stopping Rule and Alternatives . .	75
C. Optimality Criteria for Stopping Rules	81
D. Application of the Optimality Criterion . . .	86
IV. Results and Conclusions	91
A. Numerical Results	91
B. Practical Recommendations	96
C. Discussion	107
References	114
Appendix I: Crude Distributional Properties of the Weighted Least Squares Estimators of II.C	117
Appendix II: The Multivariate Integral Problem	122

CHAPTER I: A CONTEXT FOR STEEPEST ASCENT

I.A THE EXPERIMENTER'S PROBLEM

Brief descriptions of the statistical procedures known generically as response surface methodology (RSM) abound in the literature. Nearly every paper on the subject begins with at least an outline of the response surface problem, and complete descriptions are available in Myers ([19]) and Davies ([10]), Chapter 11). Rather than reproduce one of these expositions in detail, I prefer to cover the same ground subjectively, using the description of methods to illustrate one possible point from which to view recent developments in the field. Each of the methods may be thought of as contributing to efficient solution of what I shall call the "experimenter's problem" (to distinguish it from the "statistician's problem," described in I.B.). The experimenter's problem is to optimize a response y (per cent yield, for example) which is a function, usually non-linear, of several controllable concomitant variables x_1, x_2, \dots, x_k . The usual solution of the problem directs the experimenter to approximate the response function $y(\underline{x})$ by fitting a graduating polynomial $\hat{y}(\underline{x})$ to a set of observations $y(x_1), y(x_2), \dots, y(x_n)$ at designed levels of the concomitant variables. The experimenter then selects a combination of levels of the concomitant variables (representing a set of

operating conditions) which are optimum in that they maximize the graduating polynomial.¹

Contrary to the probable impression created by the preceding summary, the experimenter's problem is not given to trivial solution, because the complex nature of most response functions defies approximation by a single polynomial model over the entire range of the independent variables. Observations from one subset of the domain may lead the experimenter to believe that his function is linear, i.e., first-order, while observations from some other region would indicate that the function is second order, and observations from a third region would suggest yet another approximation. To appreciate the challenge of the experimenter's problem one might imagine a blind man trying to find the summit of a large mountain aided only by an altimeter of dubious precision. His sole recourse from trial and error would be to take systematic readings of the altitude within a limited area in an attempt to draw inferences about the local terrain, to proceed on the basis of his inferences until they were shown by new readings to be no longer true, and then to begin again with new inferences

¹Or minimize the graduating polynomial. Since the minimum of any function can be obtained as the maximum of its additive inverse, most optimization problems are stated in terms of maximization. Throughout this thesis, when the word "maximization" occurs in discussion of the experimenter's problem, the reader should understand that "minimization" could be substituted. "Optimization" will generally be reserved for discussion of the statistician's problem.

drawn from still more readings. This is precisely the strategy employed in the typical solution of a response surface problem, as illustrated below for artificial data.

EXAMPLE I.1

Stage 1: First-order estimation. The experimenter first fits the model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ from the four observations of a 2^2 factorial design using least squares.

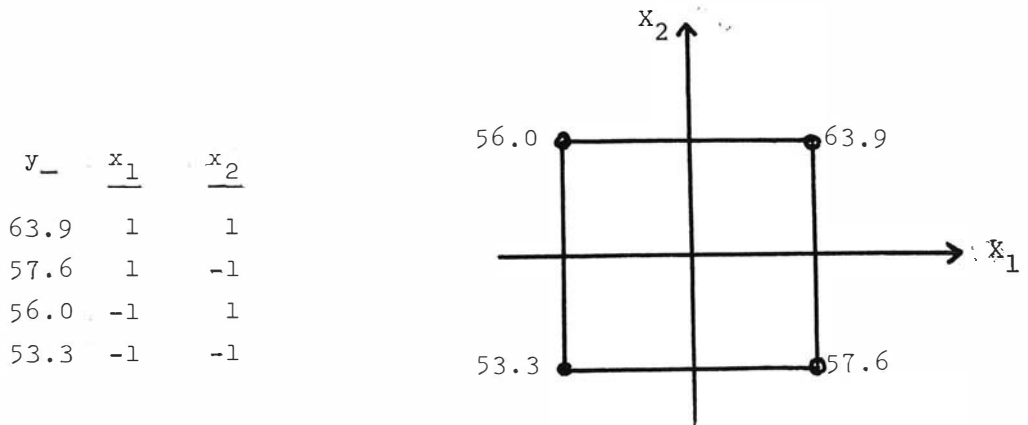


Fig. 1 First-order estimation

The fitted response function is:

$$\hat{y}(x_1, x_2) = 57.7 + 3.05x_1 + 2.25x_2$$

Stage 2: Steepest Ascent. An estimated path of steepest ascent is given parametrically as

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = t \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = t \begin{bmatrix} 3.05 \\ 2.25 \end{bmatrix} = t' \begin{bmatrix} .805 \\ .594 \end{bmatrix}$$

Observations are taken repeatedly along this line until a decrease in response is observed.

t'	x_1	x_2	y	t'	x_1	x_2	y
1	.805	.594	61.8	5	4.025	2.970	83.3
2	1.610	1.188	67.7	6	4.830	3.564	92.7
3	2.415	1.782	74.7	7	5.635	4.158	94.2
4	3.220	2.376	81.9	8	6.440	4.752	92.5

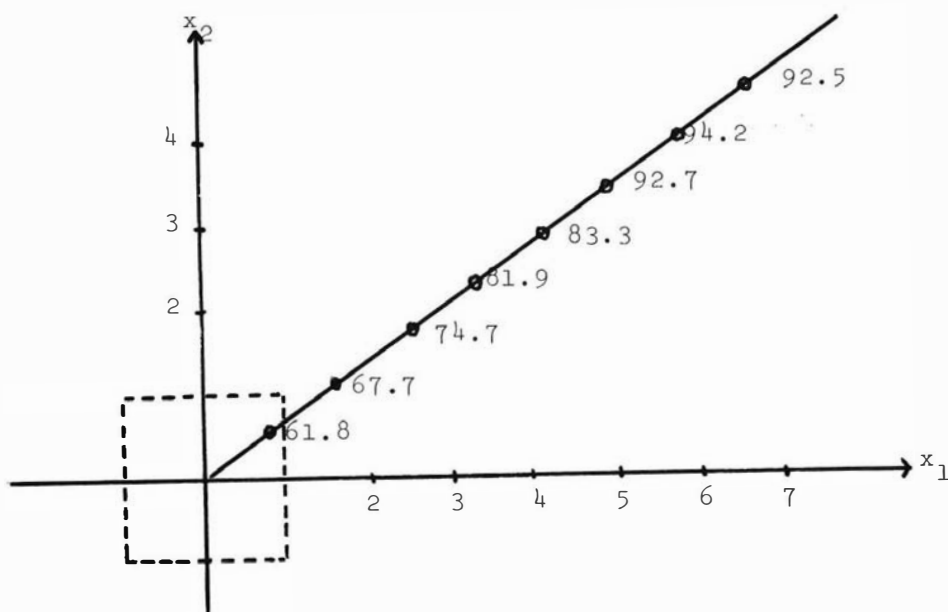


Fig. 2 Steepest Ascent

Stage 3: Second-order estimation and canonical analysis.

The experimenter now fits the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{11} x_{1i}^2 + \beta_{22} x_{2i}^2 + \beta_{12} x_{1i} x_{2i} + \epsilon_i$$

from the nine observations of a central composite design.

y	Uncoded		Coded	
	x_1	x_2	x_1^*	x_2^*
89.4	6.635	5.158	1	1
99.7	6.635	3.158	1	1
83.5	4.635	5.158	-1	1
92.9	4.635	3.158	-1	-1
84.0	5.635	5.572	0	$\sqrt{2}$
97.1	5.635	2.744	0	$-\sqrt{2}$
96.8	7.049	4.158	$\sqrt{2}$	0
87.1	4.221	4.158	$-\sqrt{2}$	0
94.2	5.635	4.158	0	0

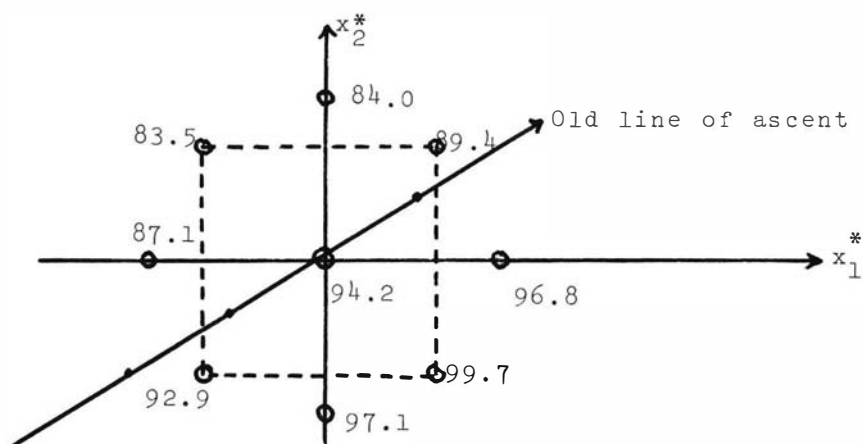


Fig. 3 Second-order estimation

The new fitted response function for the coded variables is

$$\begin{aligned}\hat{y}(x_1^*, x_2^*) &= 94.20 + 3.30x_1^* - 4.78x_2^* - 1.09x_1^{*2} - 1.79x_2^{*2} - 0.23x_1^*x_2^* \\ &= 94.20 + (x_1^*x_2^*) \begin{bmatrix} 3.30 \\ -4.78 \end{bmatrix} + (x_1^*x_2^*) \begin{bmatrix} -1.09 & -0.12 \\ -0.12 & -1.79 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix}\end{aligned}$$

which can be rewritten $\hat{y} = \alpha - \lambda_1 w_1^2 - \lambda_2 w_2^2$, where λ_1 and λ_2 are eigenvalues of the matrix of the quadratic form.

$$\lambda_1 = -1.81 \quad w_1 = .162(x_1^* - 1.67)^2 + .973(x_2^* + 1.45)^2$$

$$\lambda_2 = -1.07 \quad w_2 = .973(x_1^* - 1.67)^2 - .162(x_2^* + 1.45)^2$$

Clearly $\hat{y}(x^*)$ is maximum only when $w_1 = w_2 = 0$, so the optimum operating conditions are given by $x_1^* = 1.67$, $x_2^* = 1.45$, which, in terms of the uncoded variables, becomes $x_1 = 7.30$, $x_2 = 2.71$. The actual response function used to generate the observations was

$$\eta(x_1, x_2) = 50 + 50 \exp\{-.03(x_1 - 7)^2 - .05(x_2 - 3)^2\},$$

with a maximum of 100 at $x_1 = 7$, $x_2 = 3$.

I.B THE STATISTICIAN'S PROBLEM

The statistician who seeks to improve existing response surface methods faces a different optimization problem: that of finding the most efficient method of solving the experimenter's problem. He seeks ideally to define a procedure which, independent of the form of the true response $y(x)$, will enable the experimenter to determine his optimum operating conditions as economically as possible, i.e., with

a minimum number of observations. The statistician's optimization problem, in contrast to the experimenter's, does not lend itself to direct and systematic analysis. Whereas the experimenter's goal is clearly defined in terms of a single function y of a fixed set of controllable variables x_1, x_2, \dots, x_k , no such easy characterization is possible for the statistician's problem. Ideally he would take as his objective function (that quantity whose maximum or minimum defines the objective) the expected number of observations required by the experimental procedure to determine the optimum operating conditions. It is unfortunate but obvious that almost nothing can be said about the form of such an ideal objective function: while certain of its parameters are easily set down, e.g., the form of the response $y(\underline{x})$, the number k of its independent variables, and limits on their range, these are all characteristics of a particular experimenter's problem, and hence beyond the statistician's control. In order to determine an optimum method, the statistician must be able to write his objective function in terms of parameters which not only are subject to his control, but which at the same time completely characterize the class of methods from which he will select his optimum. In view of the clear impossibility of such a parametric characterization of all eligible response surface methods, the statistician is reduced to consideration of isolated segments of

the entire optimization problem: by making restrictive assumptions on the response function and defining new objective functions which are less intuitive but more amenable to analysis than the ideal "least number of observations," statisticians are able to state problems whose solutions yield significant increases in the efficiency of existing methods. Much of the recent progress in response surface work can be attributed to judicious selection of functions to be optimized, (or equivalently, optimality criteria to be satisfied), predicated on some natural restriction of the experimental conditions. It is within this context of restriction and optimization that I shall first outline a few of the current procedures before introducing some possibilities for further improvement which will be considered in detail in later chapters.

I.C BRIEF DESCRIPTION OF CURRENT METHODS

To the extent that tradition has had time to accumulate, response surface methodology traditionally separates into two areas, analysis and design. Analysis refers to the methods (1) of estimating the graduating polynomial \hat{y} which approximates the true response y , and (2) determining that particular combination of levels of the independent variables x_1, x_2, \dots, x_k which maximizes (or minimizes) \hat{y} ; design refers to the methods for selecting conditions under which observations are taken for use in the estimation of $E(y|x)$.

Design further separates into classes of procedures according to the restrictions placed upon the experimental conditions. The usual restrictions specify (1) some region of experimental interest, and (2) an assumed maximum order of the graduating polynomial to be fitted. The major subdivisions of response surface methodology are shown below.

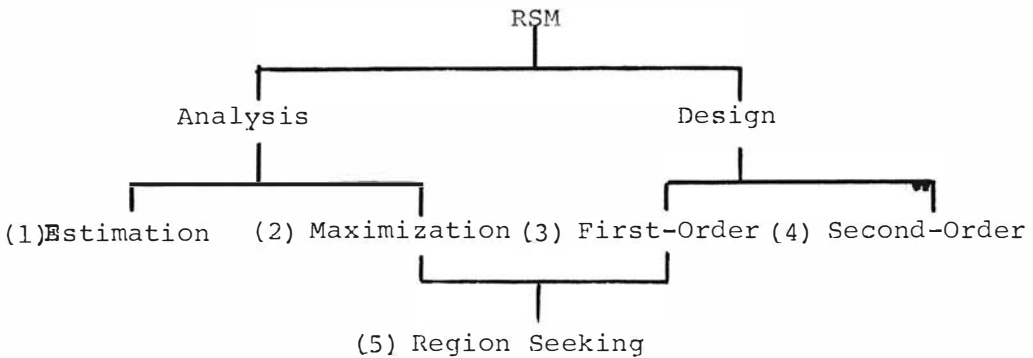


Fig. 4 Major Subdivision of RSM

Of the five numbered subdivisions in Figure 4, only the second, maximization, does not fit the pattern of restriction and optimization, and will not be discussed separately. The fifth subdivision, region seeking procedures, is the subject of detailed discussion and investigation in the following chapters; the remaining three areas are outlined below.

In reviewing the subdivisions within response surface methodology one must resist the seduction that no other partitioning of the statistician's problem could be equally appropriate or effective. Quite possibly no other partitioning would deliver sub-problems so precisely suited to traditional methods, but one must not blindly endow tradition with a propriety or effectiveness it has not earned: current methods exhibit a disparity of power which reduces their effectiveness to that of their weakest link. The strength of designed regression experiments is neutralized by the frailty of improvised region-seeking procedures, and whatever admiration the statistician commands by adapting his problem to fall within the reach of traditional methods evaporates as one discovers the imperfections in the region-seeking procedures which bind together the more established methods. Unless the region seeking procedures can be greatly improved, it may be necessary to reexamine the usual partitioning of the problem.

I.C.1. ESTIMATION

Given a set of N observations $y' = (y_1, y_2, \dots, y_N)$ on the response corresponding to points $\tilde{x}'_1, \tilde{x}'_2, \dots, \tilde{x}'_N$ in the space of concomitant variables, the experimenter postulates some form of the usual polynomial regression model $y = X\beta + \epsilon$ where ϵ is a vector of independent random variables, X is a matrix composed of the row vectors \tilde{x}' , and $\tilde{\beta}$ is a parameter vector

to be estimated. The degree of the postulated polynomial is assumed sufficient to allow "good" approximation of the response function inside some region of the independent variables; equivalently the elements of the random vector $\underline{\varepsilon}$ are assumed to have small variances, with expectations equal to zero. Under these assumptions the experimenter can justifiably anticipate that a set of operation conditions \underline{x}_0 which maximizes a fitted response $\hat{y}(\underline{x}) = \underline{x}'\hat{\underline{\beta}}$ will produce a near-maximum in the expected response $E(y|\underline{x})$, provided only that the procedure used to estimate $\underline{\beta}$ works reasonably well.

Two optimality criteria have been developed to aid the experimenter in his choice of an estimation procedure, although to date only the first has found practical application.

1) If the experimenter has confidence in his assumptions as to the form of his model, he will readily accept the usual least squares estimator $\hat{\underline{\beta}} = (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{y}$, which satisfies the optimality criterion of minimum variance when consideration is restricted to the class of linear unbiased estimators. (If the experimenter makes the further assumption that $\underline{\varepsilon} \sim N(0, \sigma^2 \underline{I})$, $\hat{\underline{\beta}}$ is then minimum variance for the class of all unbiased estimators.)

2) For the less confident experimenter, who, for example, may have postulated a second order model despite

suspicion that the true response is third order, recent authors Hader, Karson, and Manson ([15]) have introduced a method of estimation which minimizes a function of the bias in the fitted response due to the presence of higher order terms in the true model.

Despite its impracticality - the method requires a design which permits estimation of the higher order terms responsible for the bias - the proposal of Hader, et. al. merits further consideration because the restrictions it imposes on the experimental situation are more realistic than those imposed by the method of least squares. Very often the experimenter is more concerned with bias resulting from an inadequate model than with variance, and if his concern is well founded, with additional unestimated parameters γ making the true model $y = X\beta + Z\gamma + \varepsilon$ instead of $y = X\beta + \varepsilon$, the supposedly unbiased least squares estimator $\hat{y} = X\hat{\beta}$ carries a bias $[I - X(X'X)^{-1}X']Z\gamma$ ([7]). Until further work is done, however, the experimenter has no alternative to the least squares estimator. He must accept it, exploit its advantages, and attempt to circumvent its disadvantages through the only mechanism subject to his control - design of the concomitant variables.

I.C.2 DESIGN FOR FIRST-ORDER MODELS

In the early stages of an investigation, the experimenter often assumes that his region of interest is one in

which a linear (first-order) polynomial adequately describes the response function. He takes as his model

$$y(x_i) = \beta_0 + \sum_{j=1}^k \beta_j x_{ji} + \epsilon_i = \beta_0 + x_i' \beta + \epsilon_i \quad i = 1, 2, \dots, N$$

or $\underline{y} = X \underline{\beta} + \underline{\epsilon}$, and makes the distributional restrictions

$E(\underline{\epsilon}) = \underline{0}$, $\text{Var}(\underline{\epsilon}) = \sigma^2 I$. When these conditions are in force, the least squares estimator $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$ is unbiased for $\underline{\beta}$ with dispersion matrix $\text{Var}(\hat{\underline{\beta}}) = \sigma^2 (X'X)^{-1}$.

Three quantities guide the experimenter in his selection of an optimal design for fitting a first-order model: variance, bias, and number of observations.

1) Variance. Although the use of least squares estimation guarantees minimum variance regardless of the design used, the dispersion matrix $\text{Var}(\hat{\underline{\beta}}) = \sigma^2 (X'X)^{-1}$ is nevertheless a function of the matrix X and hence subject to further optimization by proper design. By writing $(X'X)^{-1}$ as a matrix of cofactors divided by $\Delta = |X'X|$, and writing the determinant Δ in a Cauchy expansion, it can be shown that the variances of the elements of $\hat{\underline{\beta}}$ are smallest when $X'X$ is diagonal, a condition which is satisfied by all orthogonal designs ([7]).

2) Bias. Although the presence of second-order terms ordinarily biases the estimates $\hat{\underline{\beta}}$, certain designs protect against this bias, or permit a test for its presence. In particular, if the model $\underline{y} = X \underline{\beta} + \underline{\epsilon}$ is used to estimate $\underline{\beta}$

when the true model is $\underline{y} = X \underline{\beta} + Z \underline{\gamma} + \underline{\varepsilon}$, the estimate $\hat{\underline{\beta}}$ will be biased by an amount $(X'X)^{-1}X'Z \underline{\gamma}$, so that if X is chosen to be orthogonal to Z , $X'Z = \Phi$, and there is no biasing ([7]).

3) Number of Observations. A first-order experiment is often part of a region-seeking procedure which will lead the experimenter away from the area where the first-order function fits well. Therefore, the experimenter's desire to minimize variance and bias with an elaborate design must be tempered by the realization that observations used in the first-order experiment rarely contribute directly to later estimates of the response function.

There is no ideal design which simultaneously minimizes variance, bias, and number of observations, but two classes of commonly used orthogonal designs allow the experimenter to choose between economy on one hand and lower variance with protection against bias on the other. The most economical designs are the simplexes, which require but one observation for each parameter to be estimated. The two dimensional simplex is represented geometrically by the vertices of an equilateral triangle, the three dimensional simplex by the vertices of a tetrahedron, and in general, the simplex design for k variables by $k + 1$ points evenly distributed on a k -dimensional hypersphere about the design center. The simplex design, though economical, has the disadvantage of

being "saturated" - every point is "used" in the estimation of parameters, and there are no observations left for independent estimation of error variance or lack of fit. In addition, the simplex results in biased estimates when the true model is quadratic.

As an alternative first-order design, the experimenter may choose a 2^k factorial or fractional factorial. These designs allow unbiased estimation of the first-order coefficients in the presence of second-order terms, and because they involve more points than the corresponding simplex designs, estimate the parameters with greater precision. When replicated observations are taken at the design center, factorial designs provide a test for lack of fit, and when lack of fit is significant, indicating that the model is second order, the factorial design can be easily augmented to form a central composite design, the most commonly used design for estimating second-order models.

I.C.3 DESIGN FOR SECOND-ORDER MODELS

Second-order designs are rarely used before the final stages of a response surface problem because they require the experimenter to invest in a large number of observations in order to estimate the $(k+1)(K+2)/2$ parameters of the second-order polynomial model:

$$y_i = \beta_0 + \sum_{j=1}^k \beta_j x_{ji} + \sum_{j=1}^k \sum_{\ell=1}^k \beta_{j\ell} x_{ji} x_{\ell i} + \varepsilon_i$$

$$= \beta_0 + \underline{x}_i' \underline{\beta} + \underline{x}_i' B \underline{x}_i + \varepsilon_i$$

where $B =$

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1k} \\ & \beta_{22} & \dots & \beta_{2k} \\ & & & \vdots \\ & & & \vdots \\ & (\text{sym.}) & & \beta_{kk} \end{bmatrix}$$

Nearly always the experimenter has limited his attention to a fixed region R , usually a hypersphere, in the space of the independent variables, either because he expects to find his optimum within R , or because it is not feasible to take observations outside R . In either case, he wants to use a design which will lead to a fitted response function $\hat{y}(\underline{x})$ which is free from bias and low in variance for $\underline{x} \in R$.

Three optimality criteria have been developed for second-order designs, with particular attention given the central composite design (cf. Example I.1). The first two criteria, efficiency and uniform information, involve only variance; the third, more recent criterion based on expected squared error incorporates both variance and bias.

1) Efficiency. Box and Wilson [[7]] defined the efficiency of one design with respect to another for estimating a particular parameter as the ratio of the variances of the estimates generated by the two designs, after adjustment for sample size and design spread. For example, if two

designs containing N_1 and N_2 points, respectively, have equal second moments,² then the efficiency of the first design with respect to the second for estimating β_1 is $E = \{N_2 \text{Var}_2(\hat{\beta}_1)\} / \{N_1 \text{Var}_1(\hat{\beta}_1)\}$. As an optimality criterion efficiency suffers from being a function of individual parameters, so that the efficiency of a design must be described by several numbers. Very often a design will estimate some parameters more efficiently and others less efficiently than an alternative design, so that an experimenter cannot rely on efficiency alone to select a unique "best" design for a given problem.

2) Uniform Information. Box and Hunter ([6])

shifted attention from the variance of individual coefficients to the variance of the fitted response function $\hat{y}(\underline{x})$. They state their criterion in terms of an information function $I(\underline{x})$ and its reciprocal, a variance function $V(\underline{x})$:

$$I(\underline{x}) = \frac{1}{V(\underline{x})} ; V(\underline{x}) = \frac{N}{\sigma^2} \text{Var}(\hat{y}(\underline{x})) = N \underline{x}' (X'X)^{-1} \underline{x},$$

²Design moments are defined as the usual moments of inertia obtained by weighting all points of the design equally. If i and j denote orthogonal dimensions of the design variables x_{ij} , then, for example,

$$[i] = \frac{1}{n} \sum_{u=1}^n x_{iu} = \bar{x}_i. \quad [j] = \frac{1}{n} \sum_{u=1}^n x_{ju} = \bar{x}_j.$$

$$\text{and } [iij] = \frac{1}{n} \sum_{u=1}^n (x_{iu} - \bar{x}_i)^2 (x_{ju} - \bar{x}_j).$$

requiring first that for any point \tilde{x} , $I(\tilde{x})$ be only a function of the distance $\rho = \sqrt{\tilde{x}'\tilde{x}}$ from \tilde{x} to the design center (rotatability), and second that, redefined as a function of distance, $I(\rho)$ be as nearly uniform as possible over the coded interval $[0,1]$. The uniform information criterion is useful to the experimenter whose primary concern is with the variance of his estimation procedures, and it has the advantage of characterizing designs which meet the criterion by a single measurement, but because an experimenter is rarely willing to ignore bias, the uniform information criterion has been superseded by a more flexible criterion.

3) Integrated expected squared error. Box and Draper ([4]) proposed as a function to be minimized by choice of design

$$J = \frac{N\Omega}{\rho^2} \int_R \{ \hat{y}(\tilde{x}) - \mathbb{E}y(\tilde{x}) \}^2 dx$$

where $\Omega^{-1} = \int_R dx$, N is the number of observations, and $\sigma^2 = \text{Var}(y_i)$. J is easily partitioned into a component V due to variance and a component B due to bias.

$$\begin{aligned} J &= V + B \\ &= \frac{N\Omega}{\sigma^2} \int_R \text{Var}(\hat{y}(\tilde{x})) dx + \frac{N}{\sigma^2} \int_R \{ \hat{y}(\tilde{x}) - \mathbb{E}y(\tilde{x}) \}^2 dx. \end{aligned}$$

Moments involving a single dimension are called "pure", moments involving two or more dimensions are called "mixed"; the order of a moment refers to the sum of the exponents in the factors. Pure second moments [ii] provide a rough measure of the spread of a design.

Box and Draper show that unless the true model is known, the experimenter will be unable to select a design which minimizes J , but they give two sets of conditions on the design moments which lead to all-bias designs (V assumed zero, B minimized by design), and all-variance designs (B assumed 0, V minimized by design). Providentially, they are able to show that under quite general conditions compromise designs exist which, though "close" to being all-bias designs, nevertheless remain nearly optimum in the presence of relatively large V . Thus, they are able to offer the experimenter unequivocal recommendations which enable him to select a unique second-order design for his experiment.

The three-stage evolution of optimality criteria for second-order designs illustrates at once the difficulty of the statistician's optimization problem, the success of the piecemeal approach which partitions the larger optimization problem into a sequence of smaller problems whose solutions derive from their eventual restatement in terms of appropriate objective functions, and the crucial nature of the functions chosen. It is the purpose of this thesis to approach the remaining area of response surface methodology, the region-seeking procedures, in the hopes of discovering objective functions which lead to improvements in existing methods.

I.D REGION SEEKING PROCEDURES

In its broadest sense a region-seeking procedure is any system of (1) observations at designed levels and (2) subsequent analysis which leads the experimenter away from a region of low response toward a region of high response. It would not be a gross oversimplification, however, to define region-seeking procedures as that class of methods which bridge the gap in response surface problems between first-order experiments and second-order experiments, as in Phase 2 of Example I.1. In practice the response to be maximized can be adequately approximated by a second-order polynomial in a design region containing the optimum, but the experimenter's first observations usually fall in some region remote from the optimum, where a linear function adequately represents the response. The goal of region-seeking procedures is to utilize the information provided by the first-order fit to determine a path which will lead the experimenter toward that region of second-order fit which contains the optimum.

It is essential for the discussion which follows that the reader be able to visualize in three dimensions those surfaces for which region seeking methods are designed. A thorough treatment would involve canonical reduction of second-order surfaces, and is beyond the scope of this summary of methods, but the essential characteristics of

such surfaces are embodied in the intuitive paradigm of a rounded mountain rising above a gradually sloping plain. In particular, the surface:

i) can be approximated by first order functions in large regions of the factor space (the sloping plain);

ii) cannot be represented over its entire domain by a single low-order polynomial; it may be concave upwards in some areas (the base of the mountain), and concave downwards elsewhere (the summit);

iii) must not contain local maxima other than a unique absolute maximum. It need not contain an absolute maximum, however, (e.g., a saddle or rising ridge instead of a mountain summit), and probably will contain ridges, which are local maxima with respect to some (but not all) of the concomitant variables.

The extent to which actual experimental conditions satisfy the preceding requirements determines their susceptibility to region-seeking procedures. If the factor space contains no large regions in which the response is essentially linear (i false), first-order procedures will be inadequate; if the function can be approximated over its entire domain by a single low-order polynomial (ii false), the maximum can be determined directly without region-seeking; and if isolated local maxima exist (iii false), they may be mistaken for the absolute maximum. Without discussing the merits of such an assumption, I shall assume in the remainder of

the thesis that the responses to be maximized do in fact conform to the three requirements. In addition, I shall assume that any region-seeking procedure takes as its origin some point in one of the planar regions in i ; if the experimenter is fortunate enough to take his initial observations in a region relatively near the optimum where first-order approximations to the response are inadequate, he can dispense with region-seeking altogether.

Proposed region-seeking procedures and improvements will be judged by their ability to cope with responses which satisfy the preceding assumptions. They should enable the experimenter to take sequential observations along a path which leads from some point of origin in the planar region to a point from which the maximum response can be determined from one or at most two second-order experiments. Because a region-seeking procedure is only the initial and not the final stage of a response surface experiment, one should not judge its success by the value of the maximum response it attains, but by the speed with which it leads the experimenter to a region where he can apply more powerful second-order procedures. Given two region-seeking procedures which lead the experimenter to a region containing the maximum, one should prefer the procedure which requires the lesser number of observations, regardless of the values of the response function at the respective stopping points of the two procedures.

I.D.1 CRUDE BEGINNINGS

The earliest and most primitive region-seeking procedure involves little more than trial and error, with the balance usually favoring the latter. Known as the one-variable-at-a-time method, it directs the experimenter to fix the levels of all but one of the independent variables, while taking observations at successively higher (or lower) levels of the remaining "free" variable as long as the response continues to increase. When the response drops the experimenter selects another free variable, fixes the remaining variables, and continues as before. After each of the variables in rotation has been taken to its apparent best level, the experimenter assumes that he is near the optimum. The path of the experimenter's observations traces the edges of a rectangular solid in the k -dimensional space of the independent variables:

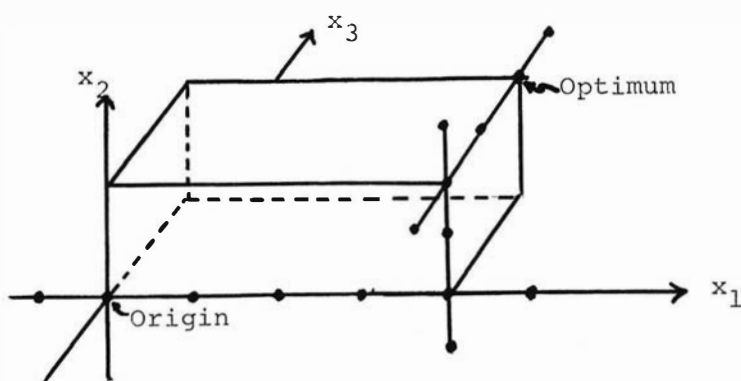


Fig. 5 The One-Variable-at-a-time Method for $k=3$

An obvious drawback of the method is that it forces the experimenter to travel a circuitous route to the apparent optimum, leaving in his wake a large number of superfluous and presumably expensive observations. The simple expedient of obtaining a preliminary first-order approximation to the response would permit the experimenter to locate subsequent observations along a path in the estimated direction of greatest increase in response. (In Figure 5, the path of greatest increase corresponds to the diagonal of the parallelepiped). As the number of independent variables increases this approach results in proportionately greater savings in terms of the number of observations required to reach the area of high response. In all but the case of two independent variables this saving alone would probably justify the use of preliminary first-order fits, but the one-variable-at-a-time method suffers from a more insidious flaw which further undermines its utility. If the error variance associated with the observations is large enough to be measured, there is an excellent chance that a decrease in observed response is due to error and does not reflect a decrease in expected response, in which case the experimenter will change free variables prematurely. Similarly for each of the other free variables there is a positive probability that the experimenter will observe a decrease in response which is premature in that it is

partially a result of random variation and does not signify a decrease in expected response. When the effects of several premature decreases are compounded, the experimenter may be led to an apparent optimum which is in fact disconcertingly far from the true optimum. Once again the proportionate loss to the experimenter increases with the number of independent variables, and once again the fitting of a first-order approximation allows the experimenter to cut that loss significantly by using some variant of steepest ascent.

I.D.2 PSEUDO-STEEPEST ASCENT BY THE SIMPLEX METHOD

The steepest ascent methods used in response surface work trace their ancestry to a family of gradient methods which are used in numerical analysis to maximize (minimize) functions whose extrema cannot be found by direct analysis ([4],p.441). In these iterative procedures, a starting point is chosen and a path away from the point in the direction of maximum increase (decrease) is determined from the partial derivatives of the function to be maximized (minimized). Successive approximations to the extremum are obtained as points along the path of steepest change. The numerical analyst is able to recompute the partial derivatives after each iteration, so that his path of steepest change receives constant revision and accurately reflects the nature of the function whose extremum he seeks. The response surface experimenter, who must pay for each new observation, cannot afford the luxury of a scheme which

would require him to run frequent first-order experiments to reestimate the partial derivatives. In the interest of economy he is forced to use some hybrid method which eschews frequent observations at the risk of lowering the rate of convergence to the optimum.

Two commonly used hybrids employ complimentary modifications of the numerical analyst's procedure. The method described in response surface literature as steepest ascent requires relatively few adjustments to the estimated path, but carries a risk that the unadjusted path will diverge from the true path before the experimenter reestimates. The alternative, known as the simplex method, allows the experimenter to reestimate the path of ascent after each observation, but restricts the location of the observations so that they form a sequence of overlapping first-order designs. Using the first method, the experimenter moves rapidly along the estimated path which he knows to be increasingly divergent from the true path; using the second method, he moves much more slowly because his observations are required to follow a broken path which is less direct than the true path, but which never diverges from the true path by much more than twice the distance between two adjacent observations.

Although the simplex method is described here as a region-seeking procedure, it was originally devised as a single-stage maximization technique, and is shown by Spendley, Hext, and Himsworth ([22]), to converge to the maximum

response even in regions where the response function cannot be approximated by a first-order polynomial. Since the simplex method can be used strictly for region-seeking, however, it merits inclusion in any comparison of region-seeking procedures.

To initiate a simplex search procedure, the experimenter takes $k+1$ observations at levels of the concomitant variables corresponding to the vertices of a regular simplex in the k -dimensional factor space. By the addition of a single observation, it is then possible to create a new simplex which shares one of its faces with the old. At each subsequent stage of the search procedure, the experimenter chooses the face of the simplex design which intersects the path of steepest ascent and completes a new simplex on that face by taking a single observation at the far vertex. The centroids of the sequentially determined simplex designs trace a path which eventually converges to the maximum response.

Spendley, et. al. show that the experimenter need not fit a polynomial to the observations in order to estimate the direction of steepest ascent because the face of a simplex which intersects the estimated path of steepest ascent will always be opposite the vertex corresponding to the lowest observation. Denoting the vertices of a regular simplex S_0 by v_1, v_2, \dots, v_{k+1} and the corresponding observations by y_1, y_2, \dots, y_{k+1} , a new simplex S_j containing the face opposite

the lowest observation y_j has vertices $\tilde{V}_1, \tilde{V}_2, \dots, \tilde{V}_{j-1}, \tilde{V}_j^*, \tilde{V}_{j+1}, \dots, \tilde{V}_{k+1}$, where

$$\tilde{V}_j^* = \frac{2}{k}(\tilde{V}_1 + \tilde{V}_2 + \dots + \tilde{V}_{j-1} + \tilde{V}_{j+1} + \dots + \tilde{V}_{k+1}) - \tilde{V}_j.$$

The authors propose three rules to regulate the application of the simplex method:

Rule 1: Ascertain the lowest reading y_p of y_1, y_2, \dots, y_{k+1} . Complete a new simplex S_p by excluding the point \tilde{V}_p corresponding to y_p and replacing it by \tilde{V}_p^* as defined above.

When the procedure is used on observations which are subject to error, there is a possibility that the system of simplexes may become anchored to some spuriously high result which is treated as if it were a genuine optimum. To reduce the risk of this we apply-

Rule 2: If a result has occurred in $(k+1)$ successive simplexes, and is not then eliminated by application of Rule 1, do not move in the direction indicated by Rule 1, or at all, but discard the result and replace it by a new observation at the same point. If the point is a genuine optimum, the repeat observation will also tend to be high. If however the result was high only by reason of errors of observation, it is unlikely that the repeat observation will also give so high a result and the point will be eliminated in due course.

Less difficulty will be caused by spuriously low results, since these will tend to be eliminated from the system fairly rapidly. However, there are advantages to be gained by also applying-

Rule 3: If y_p is the lowest reading in S_0 , and if the next observation made, y_p^* , is the lowest reading in the new simplex S_p , do not apply Rule 1 and return to S_0 from S_p . Move out of S_p by rejecting the second lowest reading (which is also the second lowest reading in S_0).

This will go some way towards reducing wandering caused by spuriously low results, but its chief purpose is that it forces the simplexes to circle continuously about an indicated optimum, rather than oscillate over a limited range. It also makes progress possible if by chance the system of simplexes should straddle a "ridge" in the factor space.

Without claiming generality for the results, one can illustrate the advantages and disadvantages of the simplex method by applying it to the function of Example I.1.

Figure 6 illustrates the path determined by the simplex method when the response function

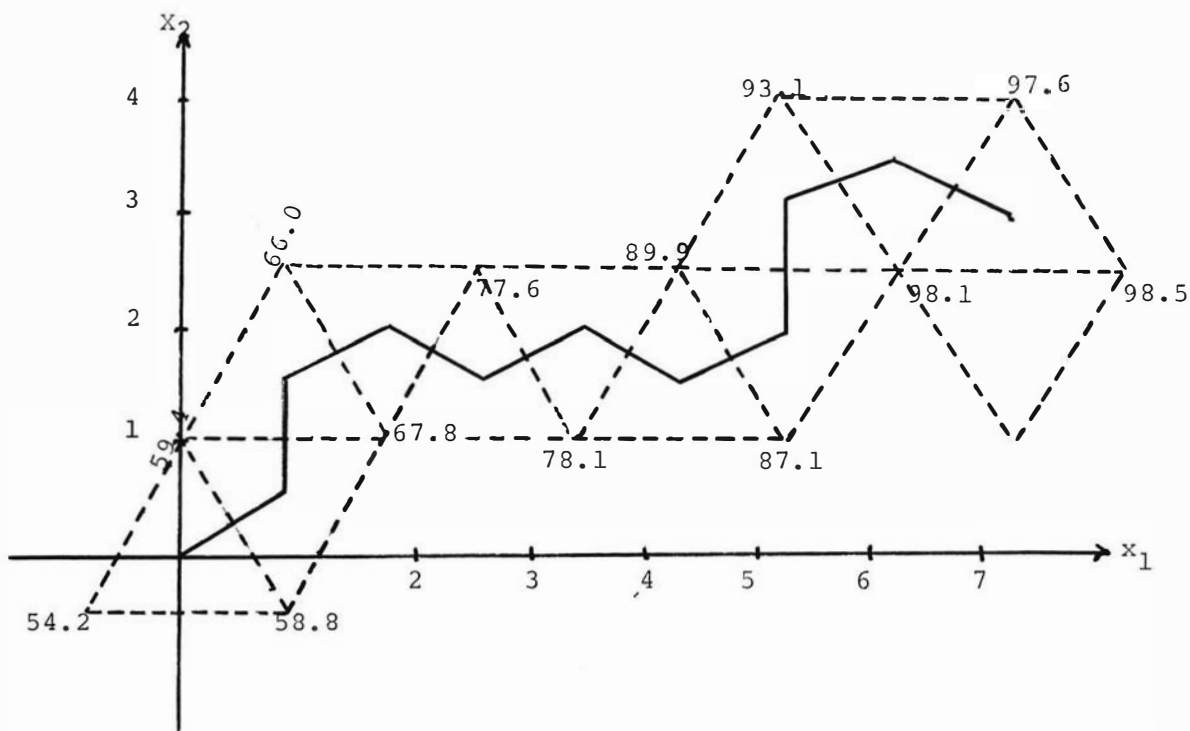


Fig. 6 The Simplex Method

$y(x_1, x_2) = 50.0 + 50.0 \exp\{-.03(x-7)^2 - .05(y-3)^2\}$ is

measured without error. The path converges rapidly to the region of near-maximum response, and illustrates both the primary strength of the simplex method - its ability to follow a non-linear path of ascent - and the method's primary weakness - its inability to follow a direct path even in that part of the factor space where the response is nearly planar.

I.D.3 TWO-PHASE STEEPEST ASCENT

The alternative to the simplex search procedure is a variant of steepest ascent which limits the experimenter to infrequent reestimation of the path of steepest ascent. Because of the resulting risk that the assumed path will eventually diverge from the true path, the experimenter must incorporate into his procedure a means of deciding when the divergence has become so significant as to justify a new determination of the path. The procedure as currently used consists of two phases, an estimation phase and a sequential decision phase. In the estimation phase, the experimenter approximates his response function using a first- or second-order polynomial model, and determines a path away from the design center in that direction which maximizes the fitted response at each point on the path. In the sequential decision phase he takes successive observations along the path, deciding after each new determination

of the response whether to continue with another observation or correct the path by refitting the response. After each refitting he tests the hypothesis that he has reached a region likely to contain the optimum he seeks.³

i) Estimation phase. After obtaining a fitted response function $\hat{y}(\underline{x})$ from a first- or second-order design, the experimenter determines the path of steepest ascent as the locus of points \underline{x}_r which maximize $\hat{y}(\underline{x}_r)$ subject to the constraint that the distance $\sqrt{\underline{x}_r' \underline{x}_r}$ from the point to the design center \underline{Q} is equal to r . Using the method of Lagrange multipliers ([7]),

$$\psi(\underline{x}_r) = \hat{y}(\underline{x}_r) - \lambda(\underline{x}_r' \underline{x}_r - r^2)$$

$$\frac{\partial}{\partial \underline{x}_r} \{\psi(\underline{x}_r)\} = \frac{\partial}{\partial \underline{x}_r} \{\hat{y}(\underline{x}_r)\} - 2\lambda \underline{x}_r$$

$$\text{So } \frac{\partial}{\partial \underline{x}_r} \{\psi(\underline{x}_r)\} = 0 \iff \underline{x}_r = \frac{1}{2\lambda} \frac{\partial}{\partial \underline{x}_r} \{\hat{y}(\underline{x}_r)\}.$$

The path is thus the locus of points whose coordinates are proportional to the first partial derivatives of the fitted response evaluated at the points themselves. If the fitted response is first-order, $\hat{y}(\underline{x}_r) = \hat{\beta}_0 + \underline{x}_r' \underline{\beta}$, and the path is determined parametrically by

³Earlier variants of steepest ascent did not contain a sequential decision phase. After estimating the path of steepest ascent from a first-order design, the experimenter immediately chose a new design center at some predetermined distance from the old center on the estimated path, ran a

$$\tilde{x}_r = \hat{\beta} / \sqrt{\tilde{\beta}' \tilde{\beta}} \quad (1)$$

If the fitted response is second-order, $\hat{y}(\tilde{x}_r) = \hat{\beta}_0 + \tilde{x}_r' \hat{\beta} + \tilde{x}_r' B \tilde{x}_r$, and the path is determined implicitly as the locus of solutions to

$$(2B - \lambda I) \tilde{x}_r = - \hat{\beta} \quad (2)$$

where the value of λ is a function of r . (The procedure based on Equation 2 is called Ridge Analysis.)

ii) Sequential decision phase. Following current practice, the experimenter takes successive observations along the path determined in i) as long as the response continues to increase. Upon observing a decrease, the experimenter usually runs a first-order design which permits a significance test on the pure quadratic coefficients of a tentative second-order model. If the second-order contribution is appreciable he augments his design to allow a

new first-order experiment, and reestimated the path ([8]). This variant minimized the risk that the experimenter's estimated path would diverge from the true path, but only at the considerable cost of an entire first-order experiment for each step taken away from the original design center.

The newer variant which includes a sequential decision phase appears to have superceded its single-phased ancestor as a region-seeking procedure and may be considered standard in a limited sense. It is reported in numerous recent papers (e.g., [3], [7], [11], [19], [22]), and though its defects are undoubtedly recognized by those who use it, no attractive alternative has appeared in the literature. When the method is described in this thesis as standard or conventional, the adjectives are used merely for identification, and should not be construed to imply universal approbation or unqualified acceptance.

second-order approximation to the response; otherwise he accepts the first-order fit and continues with steepest ascent.

Although the discussion of region-seeking has nominally recognized the existence of procedures based on second-order approximations to the response, the overriding emphasis on first-order procedures requires some justification. One could argue that any advantages accruing to the first-order methods would be further magnified if the same methods were based instead on second-order approximations, which better represent the true response over a larger portion of its domain. One could argue that a path of steepest ascent which was determined by a second-order fit would lie closer to the true path, thereby permitting the experimenter to approach closer to the optimum during the sequential decision phase before investing in a new design. One could even devise response surfaces which would sabotage first-order procedures through heavy biasing of the estimates, but which are nevertheless amenable to second-order procedures. These arguments are all impressive from the theoretical point of view, and the practitioner may indeed encounter occasional problems more suited to second-order steepest ascent, but there are two compelling practical reasons for preferring the first-order approach.

First, experience has shown that most response surfaces conform to the three conditions listed on page 21. When 1)

is in force, the response is essentially linear, and a second-order approximation is little better than a first-order approximation; and when ii) is in force, any estimates of second-order coefficients which depend heavily on the location of the observations, will not represent the response except in the immediate vicinity of the experimental region, so that several refits may be required to reach the optimum. Second, a first-order design is much more economical than a second-order design, as the following table indicates. Even the more conservative first-order design, the 2-level factorial, is approximately twice as economical as the second-order central composite design,

TABLE I: Points Required for Various Designs

Dimension	First-Order		Second-Order
	Simplex	Factorial (2^k)	Central Composite (min.)
2	3	4	9
3	4	8	15
4	5	16	25
5	6	32	43
5($\frac{1}{2}$ rep)		16	27

and the simplex affords still greater savings. For a problem involving four independent variables, the experimenter who can justify using a simplex design enjoys a 5 to 1 cost advantage over the experimenter who chooses a second-order design.

I.E. QUALITATIVE EVALUATION OF FIRST-ORDER STEEPEST ASCENT

Any attempt to compare the simplex method with its alternative, the current region-seeking variant of steepest ascent, would be almost certain to award preference to the simplex method, which was developed explicitly for response surface problems, and depends less heavily on techniques borrowed from numerical analysis. Nevertheless, because the simplex method has not been shown to satisfy any optimality criterion for region-seeking procedures, and because the alternative exhibits so many defects which might be eliminated through modification, one can only conclude that a comparison of the two procedures should be postponed until the standard variant of steepest ascent has been overhauled.

Although the standard two-phase variant of steepest ascent enjoys advantages which insure its use for region seeking in preference to other variants (the simplex method excluded), it suffers from an atavistic resemblance to its non-statistical progenitors, the numerical gradient methods. Numerical analysts operate from two assumptions which the response surface experimenter cannot accept: that only negligible error attaches to determinations of the function to be maximized, and that only the speed of their computers limits the number of such determinations. The assumptions are so fundamental that one must question the value of applying a method which incorporates both assumptions to a

problem where neither is valid. At the very least, one might have expected a more conscientious effort by statisticians to investigate the shortcomings of steepest ascent as a response surface method and to devise plugs for its more obvious loopholes.⁴ Ideally one would have asked for a method constructed de novo as the solution to some appropriate optimality problem.

Unfortunately, the ideal new method has eluded my search, and I have had to concern myself with the more mundane plugging of loopholes, of which there are distressingly many, as the following outline indicates.

1. ESTIMATION PHASE

- a) Design. Current design criteria are not appropriate for use in steepest ascent. According to current procedure, the experimenter selects his design using criteria which assume that he wants a uniformly good approximation to the response inside some spherical region R , despite his expectation that steepest ascent will lead him away from the region in which the fit is good.

⁴A notable exception exists in the form of a paper by Brooks and Mickey ([19]), in which they define and apply an optimality criterion to the selection of first-order designs for use in the older variant of steepest ascent described in the preceding footnote. Assuming the true response to be strictly first-order, and hence the true path of steepest ascent to be a straight line, they denote the angle of divergence between the true and estimated paths by θ , and seek

- b) Analysis: The method of estimation currently used to determine the path of steepest ascent obstructs the development of more efficient designs. The experimenter determines his path from estimates of the first partial derivatives at the design center, but a least squares fit of the response surface is not necessarily the best way to obtain the estimates. A more direct method of estimation would open the door to more relevant design criteria.

2. SEQUENTIAL DECISION PHASE

- a) Reestimation: Existing methods which include a sequential decision phase waste much of the information contained in the single observations taken along the path of steepest ascent. As the experimenter moves away from the original design

to maximize $W(n) = \frac{1}{n} (\cos \theta)$, where n is the number of observations required by a design. They are able to show that for a response involving k factors, $W(n)$ is maximum for simplex designs, which require the minimum of $k+1$ observations.

Unfortunately, the Brooks and Mickey criterion cannot be used as the starting point for a general evaluation of steepest ascent, because the angle θ is defined only when the true response is strictly linear, hence the criterion $\frac{1}{n} (\cos \theta)$ is meaningless when the true response is second-order. Since the goal of steepest ascent is to lead the experimenter to a region where the response is approximately second-order, any evaluation of the method must consider its behavior in the presence of second-order terms.

region R, both the variance and bias of his fitted response increase, but current procedures do not use the sequential observations to reestimate the direction of steepest ascent.

- b) Decision process: Current practice as reported in the literature reflects its numerical heritage by glossing over the importance of statistical error in deciding when to stop taking single observations and run a new first-order experiment. The existing "stopping rule" directs the experimenter to continue with single observations as long as they show an increase in response; to stop when the response decreases. As in the case of the one-variable-at-a-time method, the presence of error virtually insures that the experimenter will stop prematurely. Though no one has defended this practice, and individuals may have their own variations of the stopping rule, no one has attempted formal justification of an alternative.
- c) Design: Single observations of the sequential decision phase need not be equally spaced, and need not be taken on the line of steepest ascent; they are subject to design like any other observations. To date there is no design criterion for determining the best location for these observations.

The chapters which follow consider remedies to three of the defects listed above, although only the search for a new stopping rule (2b) proves entirely successful. Chapter II describes two of the less fruitful investigations. Section II.A attempts to solve the first-order design problem (1a) by applying the Box and Draper optimality criterion (minimum integrated squared error) to fan-shaped regions of interest leading away from the design center, but demonstrates instead that asymmetrical regions of interest can generate a computational quagmire. Section II.B tackles the problem of obtaining initial estimates for the path of steepest ascent (1b) from the observations of a 2-level factorial experiment. Since the experimenter seeks to maximize response, the design points which produce higher response are more important (they are "probably" closer to the path of steepest ascent) than the points of lower response. This logic suggests the use of weighted least squares, weighting the points of higher response more heavily in an attempt to force a better fit near the path of steepest ascent. Several examples reveal the promise of weighted least squares, but the procedure is abandoned because of its analytic intractability. Comprehensive evaluation of weighted least squares requires (in this instance) the systematic evaluation of a family of multivariate normal integrals. Coincidentally, this peripheral numerical problem returns

to haunt later chapters, and consequently merits a section of its own, II.C.

Chapter III initiates the thesis proper with a more detailed look at the sequential decision problem. A natural reparameterization reduces the general case involving k independent variables to a univariate problem involving only two parameters. The chapter concludes with the selection of an optimality criterion and outline of procedures leading to its realization. Chapter IV presents the results of the procedures and discusses practical implications.

CHAPTER II: EARLY DEAD ENDS

Both of the unsuccessful proposals examined in this chapter involve only the estimation phase of steepest ascent. In retrospect, one realizes that the estimation phase might be more resistant to improvement than the sequential decision phase simply because existing estimation procedures acknowledge the statistical nature of response surface problems whereas the sequential decision methods appear to have been invented for deterministic systems. Without benefit of this hindsight, however, one might be tempted by the recent successful work of Box and Draper to direct one's efforts toward the development of a design criterion for use in the estimation phase of steepest ascent. The most compelling need of existing criteria is to rid themselves of their dependence on restricted regions of "interest" which the experimenter must specify despite a justifiable conviction that the chosen regions bear little relevance to this problem. Both proposals attack the unnecessary region-dependence of first-order estimation procedures, but both fall short of establishing a workable optimality criterion. In the first investigation, the obvious criterion does not work, and in the second it is impossible to apply any reasonable criterion because of an analytic impasse. The first proposal attempts to exploit the region-dependence of the Box and Draper

criterion by replacing the usual symmetric (spherical or cuboidal) region with an asymmetric region elongated in the general direction of the probable path of steepest ascent; but unfortunately one finds that the Box and Draper criterion does not adapt easily to asymmetric regions. The second proposal attempts to reshape the design region ex post facto by weighting the more important observations during the estimation of parameters. This latter approach appears more successful than its predecessor, but once again the investigation breaks down before an optimality criterion can be applied.

II.A DESIGN FOR WEDGE-SHAPED REGIONS OF INTEREST

Most designs for response surface experiments owe their structure to optimality criteria established by the experimenter for a specified region of interest. For example, the usual second-order design minimizes a function of the variance and bias in the fitted response for a spherical region of the independent variables. Ordinarily the experimenter should have no difficulty in choosing his region of interest: previous experiments and external restrictions limit the area in which experimentation is likely to be fruitful. But when an experimenter wants to run a first-order experiment as part of a steepest ascent procedure he has no region of interest in the usual sense. No matter what region he selects for his design it is not a region of lasting interest because the steepest ascent procedure will

ultimately direct his attention away from the region. The experimenter might therefore hesitate to invest in a design whose optimality is defined in terms of a limited region of only temporary interest.

Unfortunately there is no reasonable design criterion which does not require the experimenter to specify a region of interest. Until such a region-free criterion appears, current methods can best be improved by adapting them to more suitable regions of interest. For example, if the experimenter had partial knowledge concerning the most plausible direction of steepest ascent, he might specify a fan-shaped region of interest with its vertex at the design center, and spreading out in the general direction of the most likely path of ascent. In two dimensions the region would be a triangle, and in three dimensions a truncated cone. A narrow angle at the vertex would correspond to a high degree of certainty regarding the direction of steepest ascent; a wide angle would correspond to a lower degree of certainty. A design could be chosen for the region using the criterion of Box and Draper: minimize the integrated expected squared error over the region of interest.

Intuitively one feels that for any reasonable region of interest the points of a design can be located so that they minimize either the variance or bias component of the Box and Draper criterion. The bias component, however, contains model parameters which cannot be eliminated from consideration

unless the region of interest is symmetric. As the following results illustrate, it is not possible to find small designs which simultaneously minimize the integrated squared bias for all possible second-order models.

Let R denote a triangular region of interest in two dimensions, bounded by the lines $x_1 = c$, $x_2 = bx_1$, and $x_2 = -bx_1$.

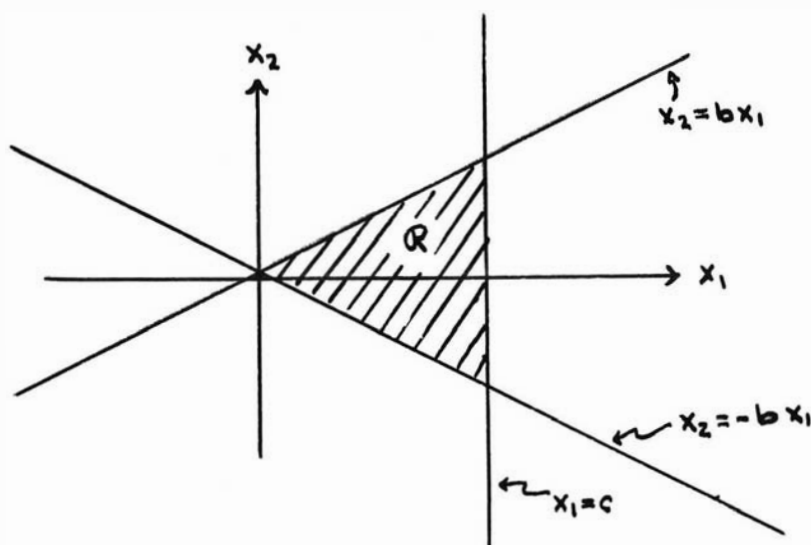


Fig. 7 Wedge-shaped region in 2 Dimensions

By performing a preliminary transformation on the design points one can obtain arbitrary values for the constants b and c ; letting $b = 1/\sqrt{3}$ and $c = 3$ will simplify future computations.

The experimenter using R as his region of interest hopes to approximate the expected response $E_Y(x_1, x_2) = \eta(x_1, x_2)$ by a linear function $\hat{y}(x_1, x_2) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

so as to minimize discrepancies between η and \hat{y} resulting from sampling error (variance) and inadequacy of the linear function (bias). In particular, he seeks a design which minimizes

$$J = \frac{N\Omega}{\sigma^2} \int_R [\hat{y}(x_1, x_2) - \eta(x_1, x_2)]^2 dx_1 dx_2,$$

where $\Omega^{-1} = \int_R dx_1 dx_2$, $\sigma^2 = \text{Var}(y)$, and N is the number of design points. Box and Draper prove the following results ([5]).

1) J can be partitioned into a variance component V and a bias component B , where

$$V = \frac{N\Omega}{\sigma^2} \int_R \text{Var}[\hat{y}(x_1, x_2)] dx_1 dx_2$$

$$B = \frac{N\Omega}{\sigma^2} \int_R [\hat{y}(x_1, x_2) - \eta(x_1, x_2)]^2 dx_1 dx_2$$

2) V and B cannot be minimized simultaneously unless η is known, because B involves parameters of the response η while V does not.

3) V alone is minimized by any design whose odd moments are all zero and whose even second moments are as large as the experimental conditions will permit.

4) If the fitted response is linear and the true response is quadratic, then a necessary and sufficient condition for minimum B is that $M_{11}^{-1} M_{12} = U_{11}^{-1} U_{12}$, where the M_{ij} are matrices of design moments and the U_{ij} are matrices of the corresponding region moments of R under a uniform distribution:

$$M_{11} = \begin{bmatrix} 1 & [1] & [2] \\ [1] & [11] & [12] \\ [2] & [12] & [22] \end{bmatrix} \quad M_{12} = \begin{bmatrix} [11] & [22] & [12] \\ [111] & [122] & [112] \\ [112] & [222] & [122] \end{bmatrix}$$

$$U_{11} = \begin{bmatrix} 1 & \mu_1 & \mu_2 \\ \mu_1 & \mu_{11} & \mu_{22} \\ \mu_2 & \mu_{12} & \mu_{22} \end{bmatrix} \quad U_{12} = \begin{bmatrix} \mu_{11} & \mu_{22} & \mu_{12} \\ \mu_{111} & \mu_{122} & \mu_{112} \\ \mu_{112} & \mu_{222} & \mu_{122} \end{bmatrix}$$

It is not difficult to find designs which satisfy the moment condition 3) above for minimum V, but condition 4) for minimum B is impossible to satisfy for small designs. For the chosen region R, with $c = 3$, $b = \frac{1}{\sqrt{3}}$, the region moments can be found trivially by integration:

$$\Omega^{-1} = \int_0^3 \int_{-1/\sqrt{3}}^{1/\sqrt{3}} dx_2 dx_1 = 3\sqrt{3}$$

$$\mu_1 = \Omega \int_0^3 \int_{-1/\sqrt{3}}^{1/\sqrt{3}} x_1 dx_2 dx_1 = 2$$

$$\mu_2 = \Omega \int_0^3 \int_{-1/\sqrt{3}}^{1/\sqrt{3}} x_2 dx_2 dx_1 = 0$$

Similarly one finds

$$\mu_{11} = \mu_{22} = \frac{1}{2}, \quad \mu_{12} = \mu_{112} = \mu_{222} = 0, \quad \mu_{111} = -\mu_{122} = -\frac{1}{5}.$$

By centering the design at (2,0) one obtains the simplification $\mu_1 = 0$.

The matrix $U_{11}^{-1} U_{12}$ of condition 4) is then

$$U_{11}^{-1} U_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{5} & \frac{1}{5} & 0 \\ 0 & 0 & \frac{1}{5} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{2}{5} & -\frac{2}{5} & 0 \\ 0 & 0 & \frac{2}{5} \end{bmatrix}. \quad 47$$

The corresponding matrix of design moments is

$$M_{11}^{-1} M_{12} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & [11][12] \\ 0 & [12][22] \end{bmatrix}^{-1} \begin{bmatrix} [11] & [22] & [22] \\ [111] & [122] & [112] \\ [112] & [222] & [122] \end{bmatrix}$$

$$= \begin{bmatrix} [11] & [22] & [12] \\ \frac{[22][111] - [12][112]}{\Delta} & \frac{[22][122] - [12][222]}{\Delta} & \frac{[22][112] - [12][122]}{\Delta} \\ \frac{[11][112] - [12][111]}{\Delta} & \frac{[11][222] - [12][122]}{\Delta} & \frac{[11][122] - [12][112]}{\Delta} \end{bmatrix}$$

Equating the top rows of $M_{11}^{-1} M_{12}$ and $U_{11}^{-1} U_{12}$ forces $[11] = [22] = \frac{1}{2}$,

$[12] = 0$, $\Delta = \frac{1}{4}$, from which it follows that $[111] = [122] = \frac{1}{5}$,

$[112] = [222] = 0$. For any given number of design points, the moment requirements reduce to a system of simultaneous polynomial equations in the coordinates of the design points.

In particular, for a design comprising the three points

(x_1, y_1) , (x_2, y_2) , (x_3, y_3) the moment requirements become

$$[1] = 0: x_1 + x_2 + x_3 = 0 \quad [2] = 0: y_1 + y_2 + y_3 = 0$$

$$[11] = \frac{1}{2}: x_1^2 + x_2^2 + x_3^2 = \frac{3}{2} \quad [22] = \frac{1}{2}: y_1^2 + y_2^2 + y_3^2 = \frac{3}{2}$$

$$[111] = -\frac{1}{5}: x_1^3 + x_2^3 + x_3^3 = -\frac{3}{5} \quad [222] = 0: y_1^3 + y_2^3 + y_3^3 = 0$$

$$[12] = 0: x_1 y_1 + x_2 y_2 + x_3 y_3 = 0$$

$$[112] = 0: x_1^2 y_1 + x_2^2 y_2 + x_3^2 y_3 = 0$$

$$[122] = \frac{1}{5}: x_1 y_1^2 + x_2 y_2^2 + x_3 y_3^2 = \frac{3}{5}$$

The three equations involving only the x_i have six common solutions given by the six permutations of (.31, .68, - .98); likewise the three equations involving only the y_i have six solutions given by the permutations of (0, $\sqrt{3}/2$, - $\sqrt{3}/2$); clearly no pair of solutions is consistent with all three equations generated by the requirements on the mixed moments. Thus no three-point design meets the Box and Draper condition 4) for minimum B.

In exactly the same way one can determine that no designs comprising 4, 5, or 6 points can be made to satisfy the equations generated by the moment requirements, so that no "small" design satisfies the Box and Draper criterion for minimum bias inside a triangular region. Larger designs might meet the criterion, but they would cost the experimenter as much as a second-order design which would eliminate bias altogether.

The failure of the Box and Draper criterion to produce optimal designs for triangular regions is not due to any defect in the criterion itself, but rather to an inconsistency inherent in its application to asymmetric regions. When the experimenter restricts himself to using small designs, he simultaneously limits his ability to minimize model inadequacies. If his region of interest is symmetric, he may be able to distribute a small number of design points throughout the region so that they minimize the bias from any possible second-order response; but if his region is not symmetric, the configuration which minimizes bias over the region for a particular second-order model may be far from optimal for some other model. The experimenter faces the choice of relaxing his restrictions on the region of interest, or making corresponding restrictions on the unknown second-order terms he seeks to protect against. The former defeats his original purpose; the latter cannot be justified from a practical point of view. The resulting stalemate has no apparent resolution.

II.B WEIGHTED LEAST SQUARES

The substantial failure of the first proposal suggests that further experimentation with design for asymmetric regions is tantamount to beating a horse whose death is all but certified. One may nevertheless retain the logic which suggested asymmetric regions by weighting the observations during estimation so as to effectively reshape the region of

the factor space in which the fitted response function best approximates the expected response function. Currently the response surface experimenter using steepest ascent determines the path along which he takes his sequential observations by fitting a linear function to the response using ordinary least squares. Given the model $y = X\beta + \varepsilon$, he selects an estimator $\hat{\beta}$ which minimizes $SS_1 = \sum (y_u - \hat{y}(x_u))^2$, where $\hat{y}(x_u) = x_u' \hat{\beta}$. If, however, he considered certain of the observations to be more "important" than others, he might weight them accordingly, choosing $\hat{\beta}$ to minimize $SS_2 = \sum W_u (y_u - \hat{y}(x_u))^2$, where W_u denotes the weight assigned the u^{th} observation. The fitted response function $\hat{y}(x)$ would then deliver a better approximation to the response near the heavily weighted points at the expense of relaxing the fit near the lightly weighted points: the experimenter sacrifices a uniformity of fit in order to improve the fit in a limited area. In particular, if he chooses to weight more heavily those points of relatively high response, he can expect a better fit in that general region of high response which is the target of steepest ascent. When second-order terms contribute significantly to the response it may be impossible to obtain a good linear approximation over the entire experimental region, so that for purposes of steepest ascent, a uniform fit is a bad fit. Weighted least squares offers the experimenter a way to insure a reasonable fit in the region of high response even

when his design criteria force him to sample with blind impartiality over a larger region of lesser interest.

The weighted least squares estimators arise from minimizing $SS_2 = \sum W_u (\hat{y}_u - \hat{y}(x_u))^2 = (\underline{y} - X\hat{\underline{\beta}})' W (\underline{y} - X\hat{\underline{\beta}})$, where W is a diagonal matrix whose non-zero elements are the weights W_u . The following results can be derived from using methods which parallel the usual least squares developments ([14]). We assume for generality, $\underline{y} = X\beta + Z\gamma$ $\text{Var}(\underline{y}) = G$, where γ is ignored in the estimation of β .

- 1) $\hat{\underline{\beta}} = (X'WX)^{-1}X'W\underline{y}$
- 2) $E\hat{\underline{\beta}} = \underline{\beta} + (X'WX)^{-1}X'WZ\gamma$
- 3) $\text{Var}(\hat{\underline{\beta}}) = (X'WX)^{-1}X'WGWX(X'WX)^{-1}$

In order to afford greater importance to the higher responses, the experimenter must choose a system of weights which is a function of the observations. In so doing he departs from the province of ordinary weighted regression by using weights which are themselves random variables. His estimate vector retains its form $\hat{\underline{\beta}} = (X'WX)^{-1}X'W\underline{y}$, but not its distributional properties, since not only \underline{y} but also W is composed of random elements. In order to evaluate the weighted estimates it seemed mandatory that at least the expectation and variance of $\hat{\underline{\beta}}$ be determined, but the complications arising from the random weighting constitute a barrier which proves impervious to analytic assault and ultimately brings the investigation to a standstill, despite the appearance of several heuristically encouraging results. Without expressions for the

expected value of the estimates it proves impossible to find an appropriate criterion for choosing a system of weights, or for evaluating the weights once they were chosen. The difficulty of judging the procedure is aggravated by the necessity for evaluating the weighted estimates in the presence of a true second-order model, so as to reflect the conditions for which the method was designed. Because no closed-form expression for $\hat{\beta}$ is available, a separate numerical determination is required for each second-order model considered, so that the comparison of estimation procedures has to occur within the context of a very limited though hopefully representative sample of second-order models. Nevertheless the results which emerge seem heartening enough to justify further investigation at some later date when the pressure of a thesis deadline is past.

II.B.1. CHOICE OF A WEIGHTING MATRIX W

In order to insure that points of higher response receive relatively greater weights, the experimenter might take his weights W_u to be monotonic functions of the observations y_u ; in particular he might choose a linear function $W_u = \gamma + \delta y_u$, where γ, δ are either predetermined parameters or functions of the responses y_u . For the sake of simplicity we shall consider weights of this form only. In matrix notation, letting $Y = \text{diag} \{y_1, y_2, \dots, y_n\}$, the diagonal

matrix formed from the observations, we write $W = \gamma I + \delta Y$:

$$\begin{aligned}\hat{\beta} &= (X'WX)^{-1}X'Wy \\ &= \frac{1}{\delta} X'[\alpha I + Y]X^{-1}\{\delta X[\alpha I + Y]y\}\end{aligned}$$

where $\alpha = \gamma/\delta$

$$= \{X'[\alpha I + Y]X^{-1}X'[\alpha I + Y]y\}.$$

In what follows each observation y_u will have weight $(\alpha + y_u)$, with the value of α to be determined. The following artificial example in one dimension illustrates the behavior of the estimation procedure as α varies. Strictly speaking there is no analog of steepest ascent in only one dimension, and it will be necessary to consider two dimensional examples, but the one dimensional example minimizes irrelevant computation while preserving several of the important characteristics of the procedure. For clarity the observations are taken equal to their expectations.

Example II.1: In the presence of a true quadratic model $\eta(x) = 3 + x + 2x^2$ we fit $\hat{y}(x) = b_0 + b_1x$ from observations at $0, \pm 1$.

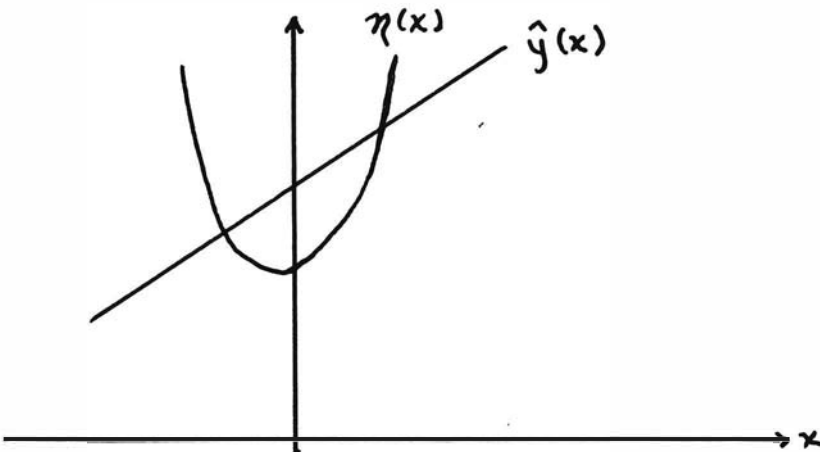


Fig. 8 One-dimensional Example

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & -1 \end{bmatrix} \quad \underline{y} = \underline{\epsilon}_y = \begin{bmatrix} 6 \\ 3 \\ 4 \end{bmatrix} \quad W = \begin{bmatrix} 6+\alpha & 0 & 0 \\ 0 & 3+\alpha & 0 \\ 0 & 0 & 4+\alpha \end{bmatrix}$$

Case 1: No weights

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X'X)^{-1}X'y = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 13 \\ 2 \end{bmatrix} = \begin{bmatrix} 4\frac{1}{3} \\ 1 \end{bmatrix}$$

Case 2: $W = Y + \alpha I$

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (X'WX)^{-1}X'W\underline{y}$$

$$= \begin{bmatrix} 13 + 3\alpha & 2 \\ 2 & 10 + 2\alpha \end{bmatrix}^{-1} \begin{bmatrix} 61 + 13\alpha \\ 20 + 2\alpha \end{bmatrix}$$

$$b_0 = \frac{26\alpha^2 - 258\alpha + 610}{6\alpha^2 + 56\alpha + 126}$$

$$b_1 = \frac{6\alpha^2 + 60\alpha + 138}{6\alpha^2 + 56\alpha + 126}$$

Plotting the linear coefficient b_1 as a function of α ,

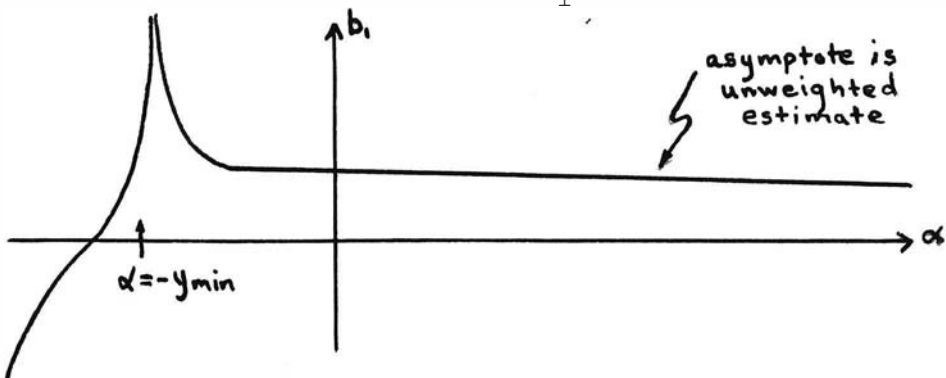


Fig. 9 Weighted estimates

Figure 9 illustrates several general properties of the weighted estimates.

- 1) Each coefficient is a rational function of α , with degree equal to $k+1$, where k is the number of independent variables. It is essential that the singularities of the rational function be avoided in choosing a value of α .
- 2) As α increases the effect of weighting gradually disappears, and the values of the weighted estimates approach the values of the unweighted estimates. This asymptotic property holds generally, for as $\alpha \rightarrow \infty$,

$$\frac{1}{\alpha} W = \frac{1}{\alpha} [\alpha I + Y] \rightarrow I \Rightarrow \hat{\beta} \rightarrow (X'X)^{-1} X'y.$$

- 3) A possible lower bound on the choice of α is given by $\alpha = -y_{\min}$. This weighting eliminates the lowest observation from the estimation procedures; lower values of α result in negative weights.
- 4) In the absence of lack-of-fit, weighting clearly has no effect.

II.B.2 TWO EXAMPLES IN TWO DIMENSIONS

In order to observe the effect of α on the direction of steepest ascent one must consider examples involving at least two independent variables. At this point one still has no criterion for selecting a value of α and hence must proceed heuristically. Primarily for lack of a clear alternative, we take $\alpha = -y_{\min}$ in the following examples,

which has the effect of eliminating the lowest observation from the design, reducing the 2^2 factorial designs used to simplexes.

Both examples assume that the true model is quadratic in two variables, and fit a linear approximation from the observations of a 2^2 experiment. As in the previous example, the observations are assumed for simplicity to be free of error.

$$\text{Model: } \eta(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

$$X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \quad W = \begin{bmatrix} y_1 + \alpha & 0 & 0 & 0 \\ 0 & y_2 + \alpha & 0 & 0 \\ 0 & 0 & y_3 + \alpha & 0 \\ 0 & 0 & 0 & y_4 + \alpha \end{bmatrix}$$

Applying the methods of Example II.1, one can obtain rational functions in α for the estimators of β_1 and β_2 . Dividing them, one obtains a rational function whose values are the slopes of the lines of steepest ascent for various values of α .

$$\hat{\beta}_2 / \hat{\beta}_1 = \frac{c_{23}\alpha^3 + c_{22}\alpha^2 + c_{21}\alpha + c_{20}}{c_{13}\alpha^3 + c_{12}\alpha^2 + c_{11}\alpha + c_{10}}$$

$$C_{13} = 16 Y_2$$

$$C_{12} = 4(4Y_{22} + Y_1Y_2 - Y_3Y_4)$$

$$C_{11} = 4(2Y_1Y_{22} - Y_2Y_{21} - Y_4Y_{23})$$

$$C_{10} = [Y_{21}(Y_3Y_4 - Y_1Y_2) + Y_{22}(Y_1^2 - Y_3^2) + Y_{23}(Y_2Y_3 - Y_1Y_4)]$$

$$C_{23} = 16 Y_3$$

$$C_{22} = 4(4Y_{23} + Y_1Y_3 - Y_2Y_4)$$

$$C_{21} = 4(2Y_1Y_{23} - Y_3Y_{21} - Y_4Y_{22})$$

$$C_{20} = [Y_{21}(Y_2Y_4 - Y_1Y_3) + Y_{22}(Y_2Y_3 - Y_1Y_4) + Y_{23}(Y_1^2 - Y_2^2)]$$

$$Y_1 = Y_1 + Y_2 + Y_3 + Y_4$$

$$Y_{21} = Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$$

$$Y_2 = Y_1 + Y_2 - Y_3 - Y_4$$

$$Y_{22} = Y_1^2 + Y_2^2 - Y_3^2 - Y_4^2$$

$$Y_3 = Y_1 - Y_2 + Y_3 - Y_4$$

$$Y_4 = Y_1 - Y_2 - Y_3 + Y_4$$

$$Y_{23} = Y_1^2 - Y_2^2 + Y_3^2 - Y_4^2$$

For each of the following examples we plot the ratio $\hat{\beta}_2/\hat{\beta}_1$ as a function of α .

Example II.2: Distant maximum

$$\text{Model: } \eta(x_1, x_2) = \beta_0' + 15\sqrt{2}x_1 - 3\sqrt{2}x_2 - 2x_1^2 + 2x_1x_2 - 2x_2^2$$

$$= \beta_0 - (w_1 - 6)^2 - 3(w_2 - 3)^2$$

$$\text{where } w_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \quad w_2 = \frac{1}{\sqrt{2}}(x_1 - x_2)$$

The response surface is a paraboloid, concave downwards, with maximum at (6.36, 2.12), approximately 6.8 units from the

design center at the origin. Figure 11 plots contours of constant response, the true line of steepest ascent, and the weighted ($\alpha = -y_{\min}$) and unweighted lines of estimated steepest ascent. Figure 10 plots $\hat{\beta}_2/\hat{\beta}_1$ as a function of α .

Example II.3: Near maximum

$$\begin{aligned} \text{Model: } \eta(x_1, x_2) &= \beta'_0 + 5\sqrt{2}x_1 - 2\sqrt{2}x_2 - 2x_1^2 + 2x_1x_2 - 2x_2^2 \\ &= \beta_0 - (w_1 - 2)^2 - 3(w_1 - 1)^2 \end{aligned}$$

$$\text{where again } w_1 = \frac{1}{\sqrt{2}}(x_1 + x_2), \quad w_2 = \frac{1}{\sqrt{2}}(x_1 - x_2).$$

Again the surface is a paraboloid; this time its maximum is at (2.12, .71), approximately 2.2 units from the design center. Figures 12 and 13 present the analogs of Figures 10 and 11.

Both examples exhibit the potential for improvement available through the use of weighted least squares, but neither provides more than circumstantial evidence that the procedure will work in general. Any of an infinite number of second-order models could have generated the same four observations used in either of the examples; any of the infinite number would then have led the experimenter to the same estimated path of ascent. If weighted estimation is to be judged superior to the usual procedure, it must demonstrate its superiority with respect to not one but all of the second-order models.

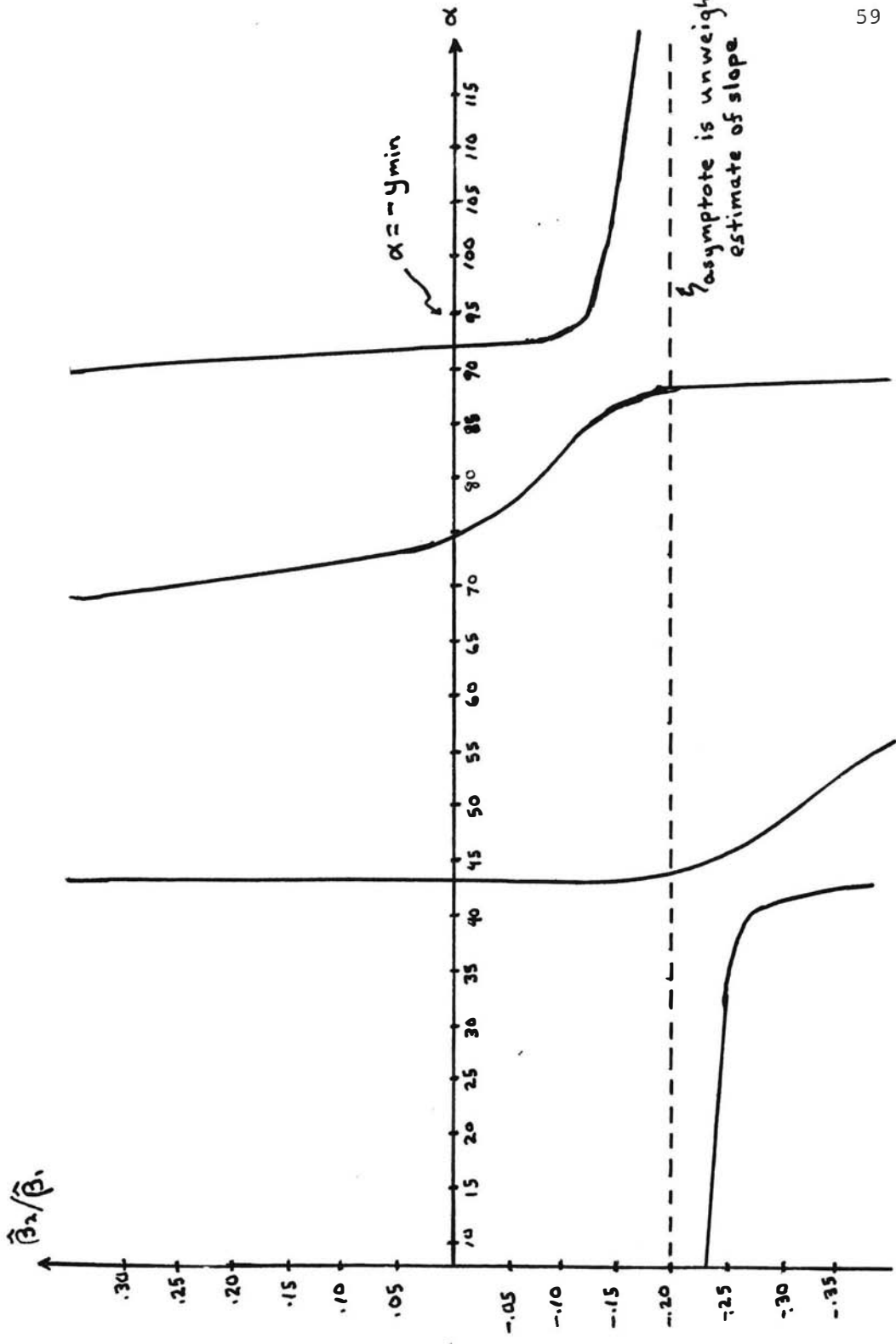


Fig. 10 Slope as a function of α : Example II2

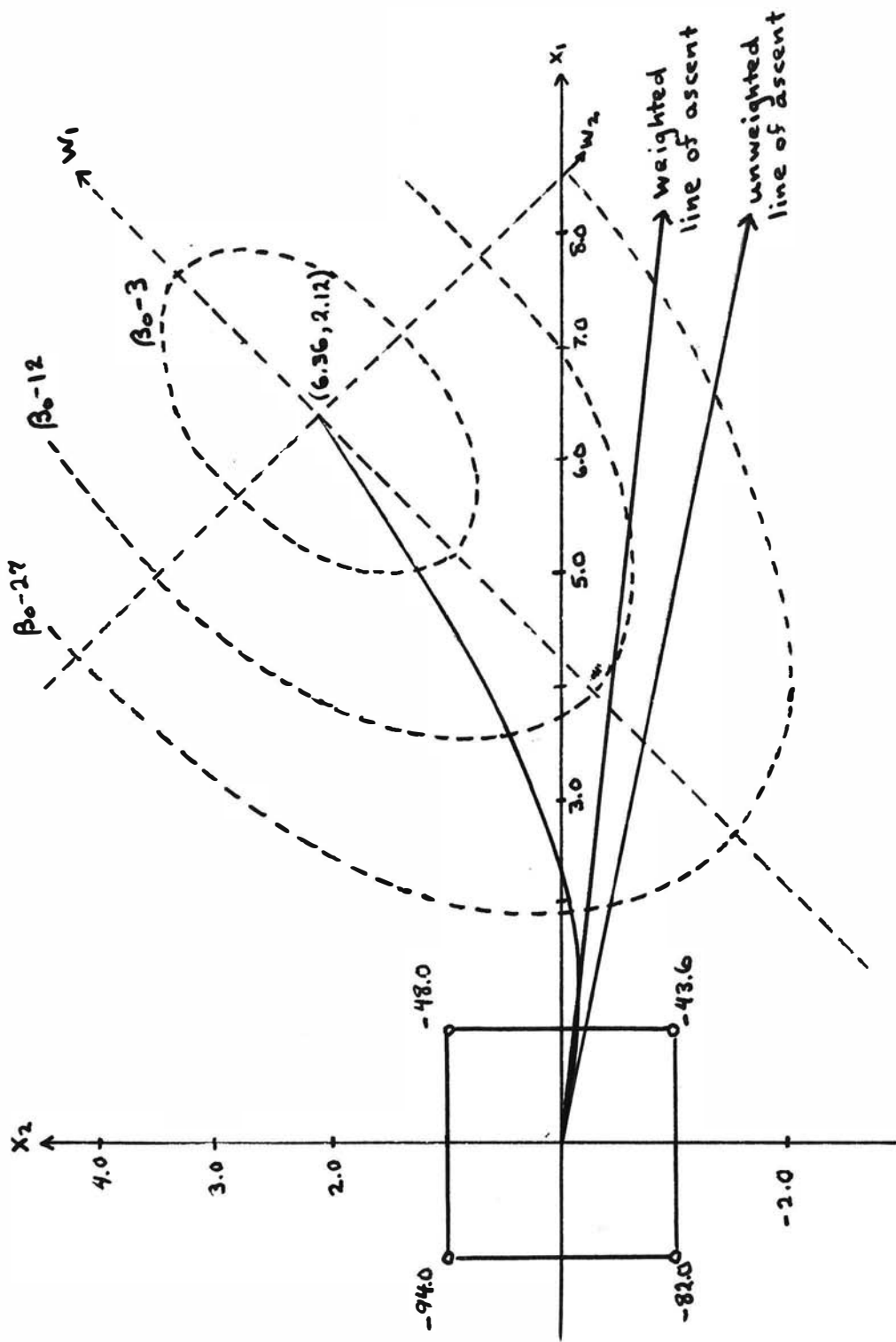


Fig. 11 Weighted ($\alpha = -Y_{\min}$) and Unweighted lines of ascent: Example II.2

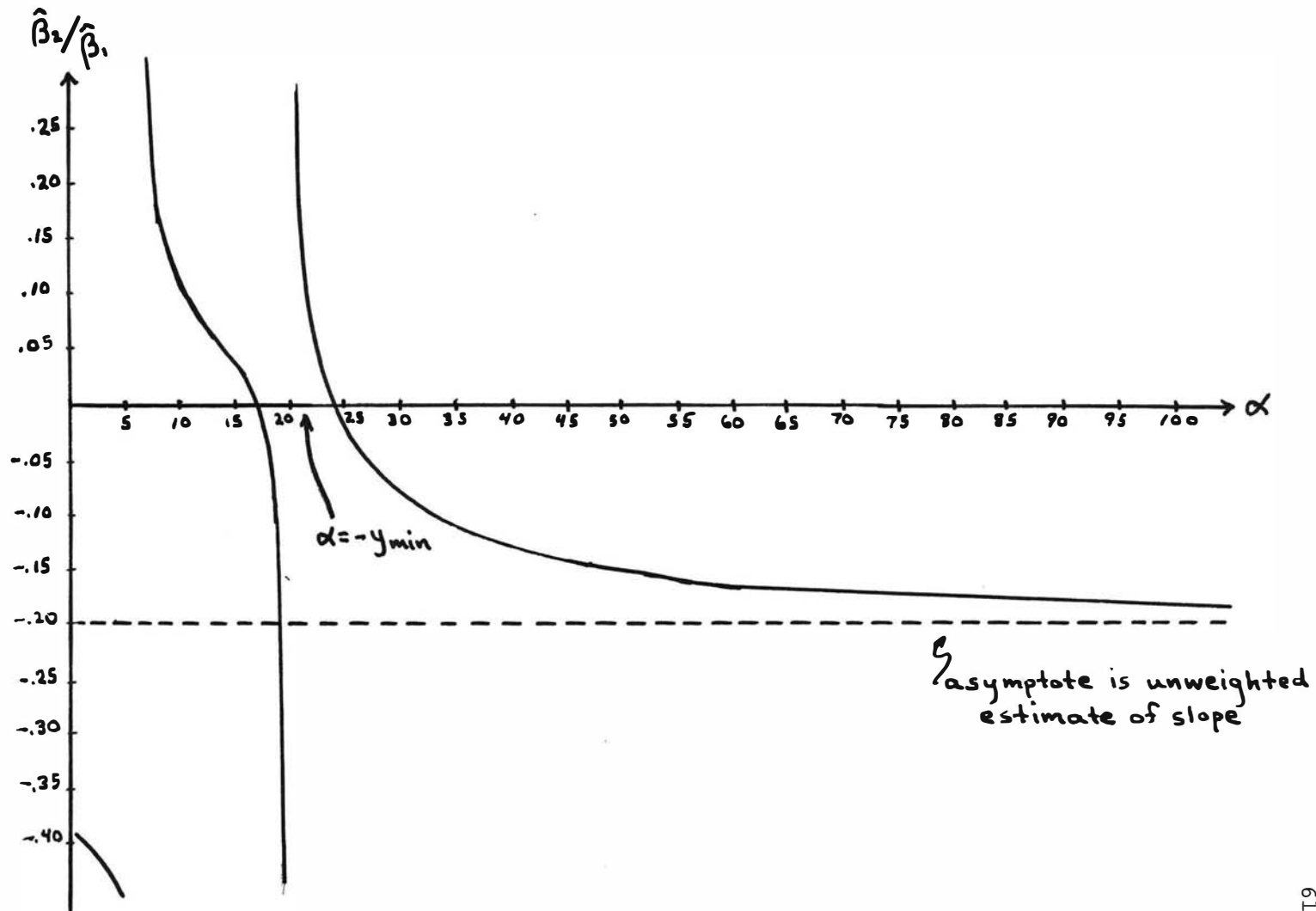


Fig. 12 Slope as a function of α : Example II.3

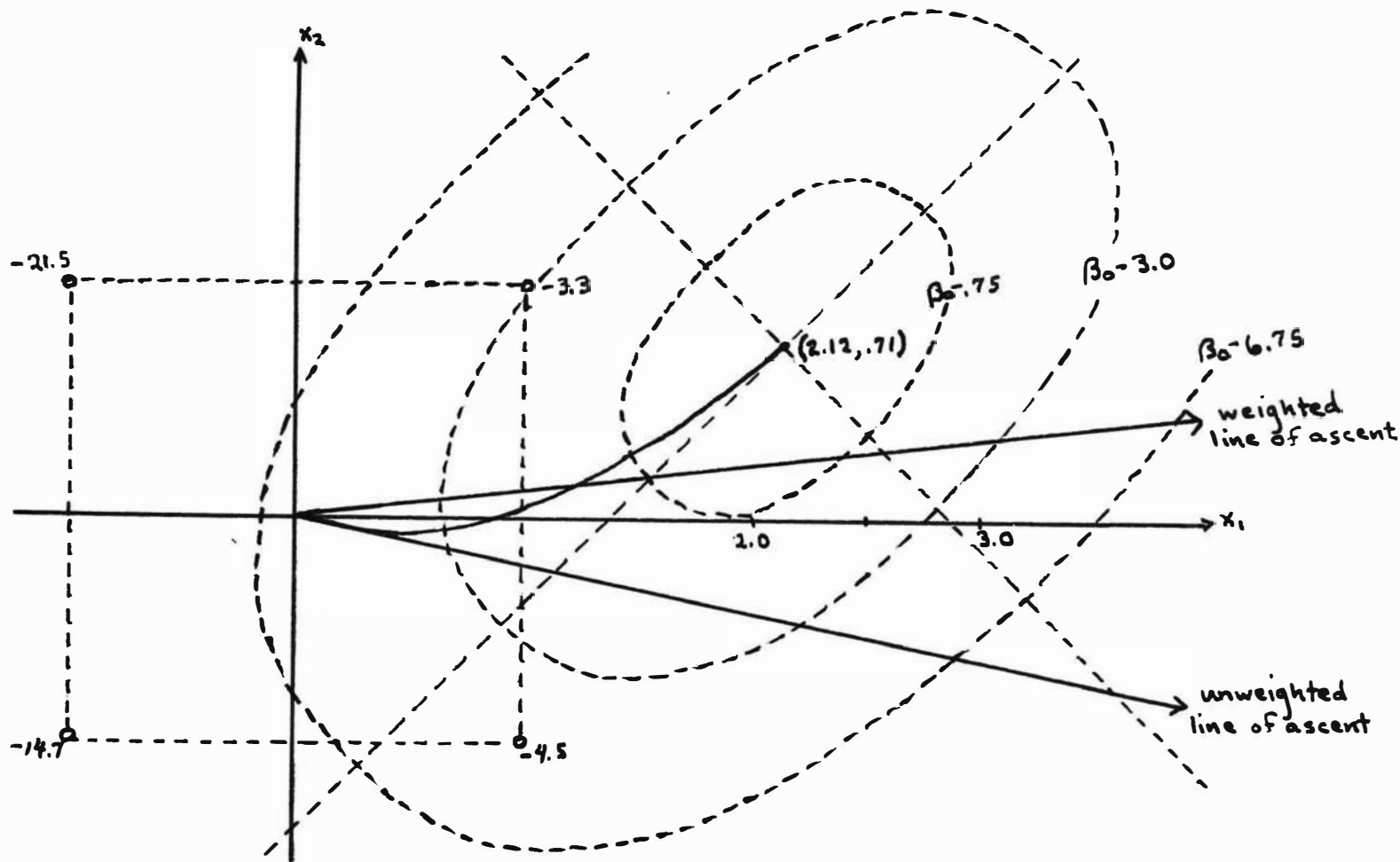


Fig. 13 Weighted ($\alpha = -y_{\min}$) and unweighted lines of ascent: Example II.3

II.B.3 GENERALIZING THE TWO EXAMPLES

The four observations of a 2^2 experiment suffice to determine estimates in the second-order model of the parameters $\beta_1, \beta_2, \beta_{12}$ but not of $\beta_0, \beta_{11}, \beta_{22}$. The normal equation corresponding to the remaining degree of freedom after estimation of $\beta_1, \beta_2, \beta_{12}$ may be written $\hat{\beta}_0 = \bar{y} - \hat{\beta}_{11} - \hat{\beta}_{22}$. One may therefore think of $\hat{\beta}_0$ as being determined by the two remaining parameters $\hat{\beta}_{11}, \hat{\beta}_{22}$, which index an infinite family of second-order models consistent with the four observations of the experiment. Any evaluation of a particular estimated line of ascent must contrive to relate its performance to the entire family of second-order models. In particular, each of the examples in II.B.2 may be generalized by arbitrary variation of the indeterminate parameters β_{11}, β_{22} .

Rather than working directly with β_{11} and β_{22} , one can obtain a reparameterization which is intuitively more satisfying. Each second-order surface $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$ possesses a unique "stationary point" at which the two partial derivatives $\partial \eta / \partial x_1$ and $\partial \eta / \partial x_2$ simultaneously vanish. Differentiating and setting the derivatives equal to zero determines the coordinates of the stationary point as the solutions to a pair of equations:

$$\begin{bmatrix} \beta_{11} & \frac{1}{2}\beta_{12} \\ \frac{1}{2}\beta_{12} & \beta_{22} \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} = - \frac{1}{2} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} .$$

Replacing the estimable parameters by their estimates b_{12} , b_1 , b_2 , one obtains a one-one correspondence between the free parameters β_{11} , β_{22} and the coordinates (x_{01}, x_{02}) of the stationary point:

$$\begin{bmatrix} \beta_{11} & \frac{1}{2}b_{12} \\ \frac{1}{2}b_{12} & \beta_{22} \end{bmatrix} \begin{bmatrix} x_{01} \\ x_{02} \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

$$\Leftrightarrow \begin{pmatrix} \beta_{11} \\ \beta_{12} \end{pmatrix} = -\frac{1}{2} \begin{pmatrix} [b_2 + x_{02}b_{12}]/x_{01} \\ [b_1 + x_{01}b_{12}]/x_{02} \end{pmatrix}.$$

One may therefore use the "parameters" x_{01} , x_{02} to index the family of second-order models consistent with a particular experiment.

Figures 14 and 15 compare the weighted ($\alpha = -y_{\min}$) and unweighted lines of ascent for various second-order models consistent with the data used in the Examples II.2 and II.3. The various stationary points are located by polar coordinates with the radius arbitrarily fixed at 3 and the angle of inclination varied by 30° increments from 0° to 330° . Overall the weighted procedure fares better than the unweighted, particularly for the data of Example II.2. Even in its present far-from-optimal state the new procedure seems to lead to an improvement over the usual methods in the majority of cases

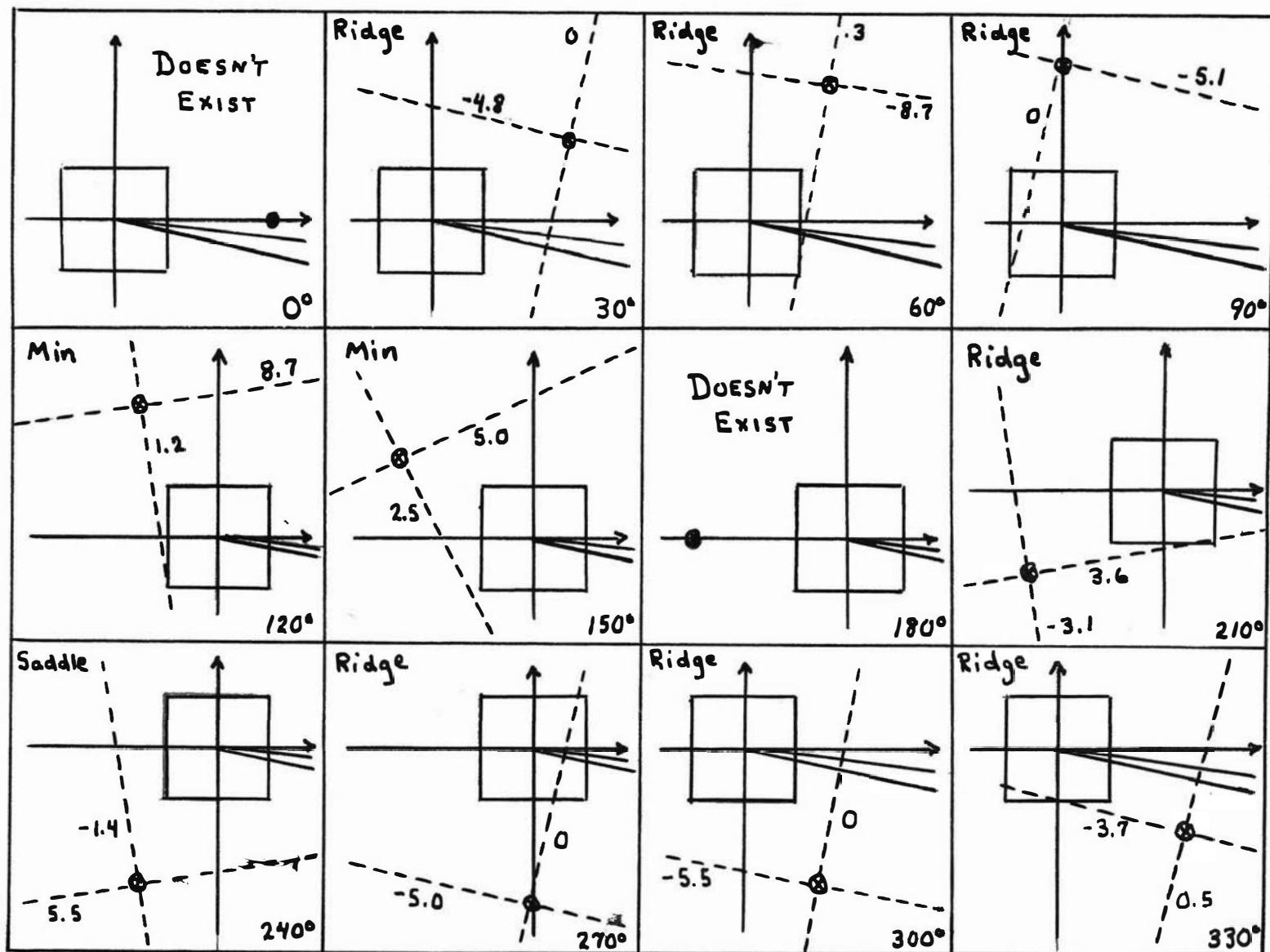


Fig. 13 Example II.2 generalized

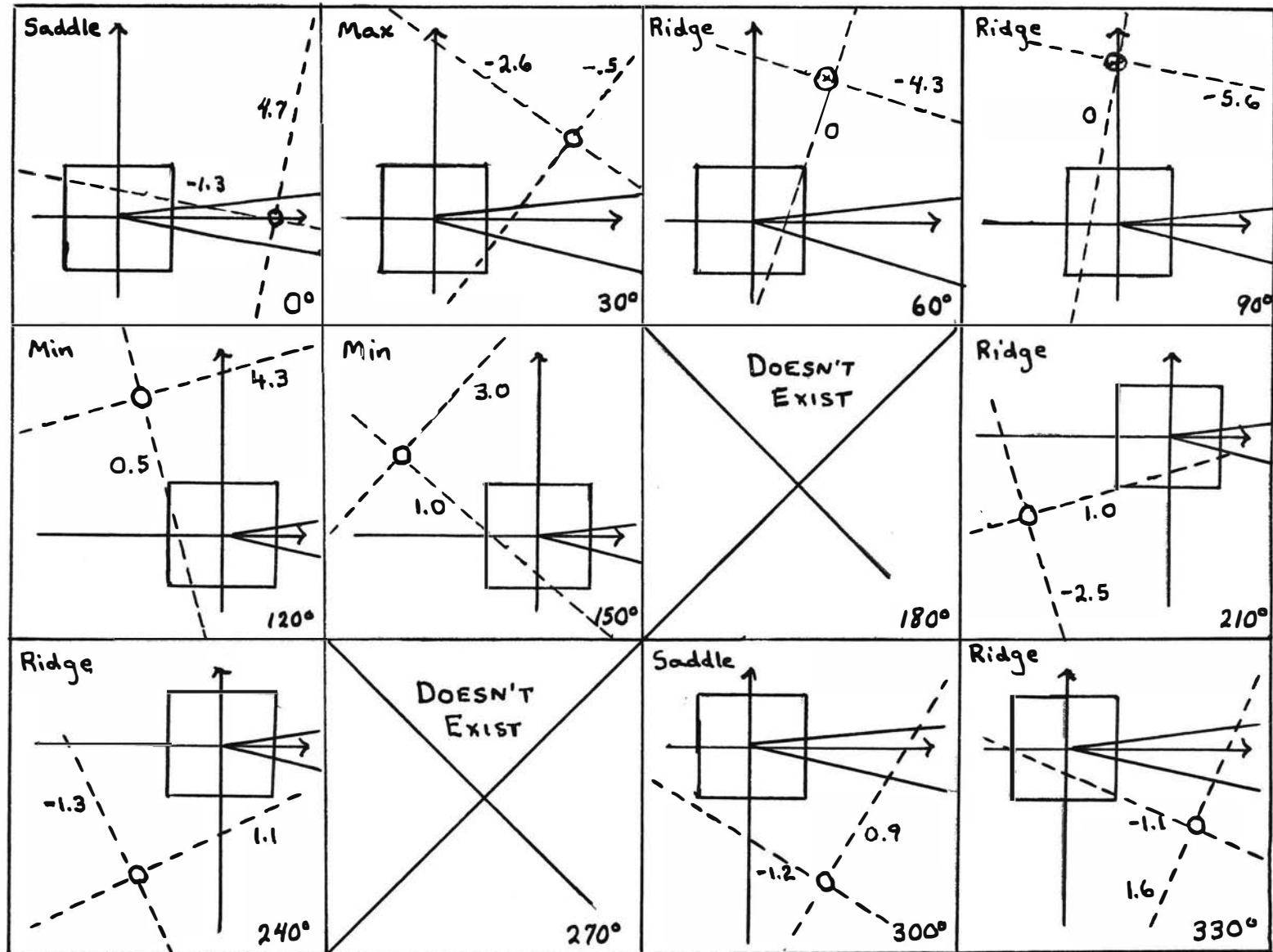


Fig. 14 Example II.3 generalized

considered; when the weighted procedure is worse than the unweighted it is a case of comparing two disasters: both work so poorly that it doesn't matter which is worse.

The behavior of the estimator in the presence of observational error does not seem to alter the above results, and so its discussion has been relegated to an appendix.

II.B.4 CONCLUSIONS AND LACK THEREOF

Reviewing the rather limited evidence of the preceding pages, one can only conclude that the case for weighted regression remains to be proved. On the positive side one may count the performance of the weighted estimator in the examples considered and the intuitive appeal of a method which is not borrowed from a peripheral discipline but designed specifically for the task at hand. On the negative side one must emphasize the limited evidence and the obstreperous secretiveness of the distributional properties which prohibits a more theoretical approach. Even though the coyness of any particular weighted estimator might be overcome by a persistent investigator with computer enough and time, there would still remain the question, "Which weighted estimator?" for the choice of a weighting scheme will ultimately determine the success or failure of the procedure. Without at least a minor analytic breakthrough it may prove impossible to find an optimality criterion which will lead to a "best".

weighting scheme.

II.C THE MULTIVARIATE INTEGRAL PROBLEM

Because the weighted estimators of II.C ultimately drowned in their own computational juices, one feels post mortem that a simpler weighting scheme might have produced a more viable estimator. The use of linear functions $w_u = \alpha + y_u$ produced weights which follow a normal distribution with continuous and infinite range. If instead the experimenter had chosen his weights from a predetermined finite set of constants, with only the assignment of the weights dependent on the observations, the weights w_u would then be discrete random variables with finite ranges, and the analysis of the weighted estimators would become much simpler. For example, for a 2^2 factorial, the experimenter might assign a weight of 8 to the largest observation, 4 to the second largest, 2 to the third largest, and 1 to the smallest. There would then be only one weighting for each possible ordering of the four observations, 16 in all. The weighting matrix W would be constant for each of the 16 orderings, so that in each case one could obtain a closed-form formula for $\hat{\beta}$ as a function of the observations. In particular for $y_1 > y_2 > y_3 > y_4$, $W = \text{diag} \{8, 4, 2, 1\}$ and

$$\hat{\beta} = (X'WX)^{-1} X'WY$$

$$= \begin{bmatrix} 15 & 9 & 5 \\ 9 & 15 & 3 \\ 5 & 3 & 15 \end{bmatrix}^{-1} \begin{bmatrix} 8_{y1} + 4_{y2} + 2_{y3} + y_4 \\ 8_{y1} + 4_{y2} - 2_{y3} - y_4 \\ 8_{y1} - 4_{y2} + 2_{y3} - y_4 \end{bmatrix}.$$

One could then compute the conditional expectation

$E(\hat{\beta} \mid y_1 > y_2 > y_3 > y_4)$, and the unconditional expectation $E\hat{\beta}$

could be obtained as the sum of all sixteen such conditional

expectations, each weighted by the probability of the

associated ordering of the observations, e.g. $\Pr \{y_1 > y_2 > y_3 > y_4\}$.

Enter the multivariate integral problem: evaluating

$\Pr \{y_1 > y_2 > y_3 > y_4\}$, which is so easy when the y_i are i.i.d.,

is far from trivial otherwise. Under the assumed model,

$\underline{y} \sim N_4(\underline{\eta}, \sigma^2 I)$, where $\underline{\eta} = X\beta$. From the joint density of

(y_1, y_2, y_3, y_4) one can write

$$P(y_1 > y_2 > y_3 > y_4) = (2\pi\sigma^2)^{-2} \int_{-\infty}^{\infty} \int_{-\infty}^{y_1} \int_{-\infty}^{y_2} \int_{-\infty}^{y_3} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^4 (y_i - \eta_i)^2} dy$$

which cannot be evaluated directly unless $\eta_1 = \eta_2 = \eta_3 = \eta_4$.

Methods of solving the multivariate integral problem in this

and other guises occupy Appendix II.

CHAPTER III: THE STOPPING RULE PROBLEM

No matter how carefully the experimenter plans and executes the estimation phase of a steepest ascent procedure, he must enter the sequential decision phase with judicious cynicism, for sooner or later his estimated path of ascent will diverge from the true path, and continued observation on the estimated path will yield a decrease in response. The efficiency of steepest ascent as a region-seeking procedure is determined by the experimenter's ability to recognize the point at which his estimate of the direction of ascent ought to be revised by refitting the model from a new experiment. If, as is currently the case, he consistently stops short of the ideal stopping point, he uses more first-order experiments than are needed to reach the optimum; if he goes beyond the ideal point before refitting, he will have to double back over the same ground at the cost of additional observations; either way, the efficiency of his procedure suffers. As long as the experimenter's observations remain in a region where the response is approximately linear, any divergence between the estimated and true paths of steepest ascent will be relatively small, and the experimenter should continue to take single observations along his estimated path. Only after the gradually increasing curvature of the response has introduced small but unestimated quadratic terms into the model does the experimenter encounter the Stopping Rule Problem: At what point does the inadequacy of his model and

the bias of his estimates proscribe further use of the estimated path of steepest ascent? Where is the ideal point at which to stop taking single observations? How can the experimenter determine its location most efficiently?

Current procedure instructs the experimenter to compare successive single observations until he observes a decrease in response. In a large number of cases this procedure stops the experimenter prematurely, causing him to invest in a new design before he needs it. The remaining two chapters seek a precise measurement of the resulting loss to the experimenter and propose an alternative stopping rule which is chosen to minimize that loss.

III.A REPARAMETERIZATION AND PROBLEM DEFINITION

In the usual steepest ascent procedure, the experimenter uses only first-order approximations to the response function, and as long as he remains in a region of the factor space remote from the optimum, first-order estimates of the path of steepest ascent serve him well. He need not consider stopping his single observations until he has reached a region where, despite his working assumption of a first-order model, the response is better represented by a second-order model:

$$\begin{aligned}
 y(x_1, x_2, \dots, x_k) &= \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^k \sum_{j=1}^i \beta_{ij} x_i x_j + \epsilon \\
 &= \beta_0 + \underline{x}' \underline{\beta} + \underline{x}' \underline{B} \underline{x} + \epsilon
 \end{aligned}$$

where $i, \epsilon \sim N(0, \sigma^2)$

$$\text{ii) } \underline{\beta}' = (\beta_1, \beta_2, \dots, \beta_k)$$

$$\text{iii) } B = \{b_{ij}\}, \text{ with } b_{ii} = \beta_{ii}, b_{ij} = \frac{1}{2}\beta_{ij} \text{ for } i \neq j.$$

72

Because it is the presence of second-order terms in the response which renders the first-order estimates inadequate and creates the need for a stopping rule, any attempt to evaluate and improve the current rule must work from the assumption that the response function is in fact second-order.

The response function $y(\underline{x})$ takes as its domain the entire k -dimensional factor space, but the domain of immediate interest to the experimenter is a one-dimensional subspace corresponding to his estimated line of steepest ascent. By restricting the domain of y to this line and reparameterizing, one can reduce the stopping rule problem to a single dimension. To achieve this reduction, let $\underline{x}(t)$ denote a point on the linear estimated path at a distance t from the origin (design center). Then the path can be written as the locus of points $\underline{x}(t) = t \underline{\lambda}$, where $\underline{\lambda}' = (\lambda_1, \lambda_2, \dots, \lambda_k)$ is the normalized vector of previously estimated first-order coefficients. Conditional on the estimated direction, the expected value $\eta(t)$ of an observation $y(t) = y(\underline{x}(t))$ taken on the path of ascent is

$$\begin{aligned} \eta(t) &= \beta_0 + (t\underline{\lambda})'\underline{\beta} + (t\underline{\lambda})'B(t\underline{\lambda}) \\ &= \beta_0 + t(\underline{\lambda}'\underline{\beta}) + t^2(\underline{\lambda}'B\underline{\lambda}) = \gamma_0 + \gamma_1 t + \gamma_2 t^2; \end{aligned}$$

and conditional upon the estimated direction $\underline{\lambda}$, observations taken on the path retain their normality and independence.

Under the usual conditions of a response surface experiment, γ_1 will be positive and γ_2 will be negative but small in absolute value. (If $|\gamma_2|$ were large, the experimenter would already be in a region of high curvature, presumably near the optimum; if γ_2 were positive, the expected response would be an increasing function of t , with no need for a stopping rule; γ_1 is positive by choice of the path of ascent.) As a function of t , the expected response will increase more and more slowly, reach a maximum, then decrease more and more rapidly (Figure 16).

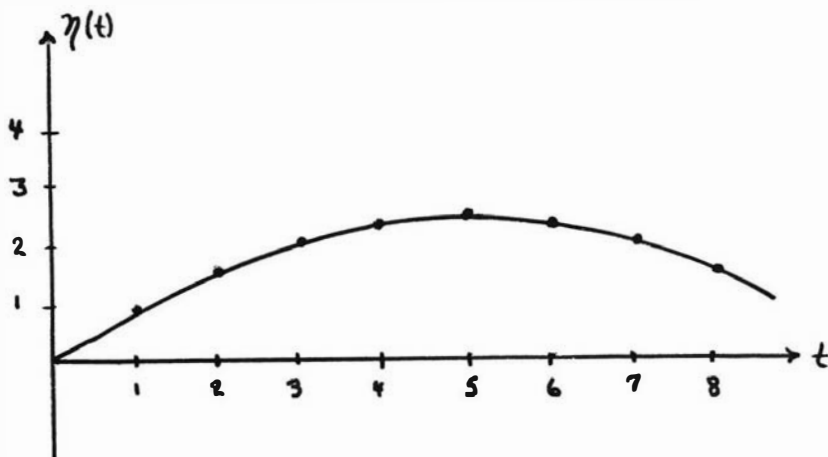


Fig. 16 Expected response for points on the estimated line of steepest ascent
($\gamma_1 = 1.0$, $\gamma_2 = -0.1$)

The logical ideal stopping point t^* is at the maximum, which lies on a ridge of the response surface. If the experimenter had some means of stopping on the ridge, he could run a new experiment to determine its orientation, then follow it to the region of optimum response.

We shall assume for simplicity that the experimenter takes his observations only at positive integral values of t , and denote these observations individually by y_j , $j=1,2,\dots$, and collectively by the vector \underline{y} of finite but possibly unspecified length. Further, we shall assume $\text{Var}(y_j)=1$. If, contrary to these assumptions, the distance between successive observations is $\delta \neq 1$, and $\text{Var}(y_j)=\sigma^2 \neq 1$, the parameters $\gamma_0, \gamma_1, \gamma_2$ should be replaced by $\sigma\gamma_0, \sigma\gamma_1/\delta, \sigma\gamma_2/\delta^2$ respectively, and the transformed responses will meet the assumptions:

$$y_j \sim N(\sigma\gamma_0 + (\frac{\sigma\gamma_1}{\delta})(\delta j) + (\frac{\sigma\gamma_2}{\delta^2})(\delta j)^2, \sigma^2)$$

$$\Rightarrow y_j/\sigma \sim N(\gamma_0 + \gamma_1 j + \gamma_2 j^2, 1).$$

In view of the simplifying assumptions the Stopping Rule Problem may be restated as follows: Given a sequence of independent normal observations $\underline{y}' = (y_1, y_2, \dots, y_j)$ for which

$$E(y_j) = \gamma_0 + \gamma_1 j + \gamma_2 j^2, \quad \text{Var}(y_j) = 1, \quad j=1,2,\dots,n$$

how can one test $H_0: E(y_n) = \max_t (\gamma_0 + \gamma_1 t + \gamma_2 t^2)$

"as efficiently as possible?" Clearly any solution to the problem will require an explicit definition of efficiency which reflects the sequential nature of the testing procedure.

Rather than present the chosen definition deus ex machina, we shall approach it indirectly by first examining the defects of the existing stopping rule and the possible advantages of some intuitively reasonable alternatives.

III.B THE CURRENT STOPPING RULE AND ALTERNATIVES

Although the current stopping rule is more a primitive rule of thumb than a rigorous procedure for testing statistical hypotheses, it can be formally restated so that the usual measurements (e.g., power) associated with hypothesis testing may be used in its evaluation. As the experimenter takes each single observation y_n on the path of steepest ascent and decides whether to take still another single observation or not, he is informally testing the hypothesis that he has not yet reached the ideal stopping point t^* : he continues if he accepts the hypothesis, stops if he rejects. He bases his test on the two most recent observations y_n and y_{n-1} : if $y_n \leq y_{n-1}$ he accepts the hypothesis; otherwise he rejects. Although he would probably state the hypothesis under test in terms of the maximum expected response $\eta(t)$, i.e., $H_0: n < t^*$, where $\eta(t^*) = \max_t \eta(t)$, his procedure simultaneously tests an equivalent hypothesis on the slope $\partial \eta(t) / \partial t$ of the expected response: since $n < t^*$ if and only if the slope of $\eta(t)$ is positive, the equivalent hypothesis is

$$H_0: \left. \frac{\partial \eta(t)}{\partial t} \right|_{t=n} > 0.$$

This restatement of the hypothesis allows one to think of the test statistic $(y_n - y_{n-1})$ as an unbiased estimator of the slope of the response function $\eta(t)$ evaluated at $t = n - \frac{1}{2}$.

$$\frac{\partial}{\partial t} \eta(t) = \gamma_1 + 2t\gamma_2$$

$$\Rightarrow E(y_n - y_{n-1}) = \gamma_1 + (2n-1)\gamma_2 = \left. \frac{\partial \eta(t)}{\partial t} \right|_{t=\frac{2n-1}{2}}.$$

Thus at each step the experimenter tests

$$H_{0n}: \left. \frac{\partial \eta(t)}{\partial t} \right|_{t = \frac{2n-1}{2}} > 0$$

using the test statistic $(y_n - y_{n-1})$, which is known to have a normal distribution with variance 2.0 and expectation as given above; one can easily determine the level α and power $1 - \beta$ of the test.

$$\alpha = \Pr\{y_n < y_{n-1} \mid E(y_n - y_{n-1}) = 0\} = .50$$

$$\beta = \Pr\{y_n \geq y_{n-1} \mid E(y_n - y_{n-1}) = \xi\} = \frac{1}{2\sqrt{\pi}} \int_0^{\infty} e^{-\frac{1}{4}(x-\xi)^2} dx$$

A plot of the power curve provides all the evidence needed to condemn the current stopping rule to a deserved ignominy.

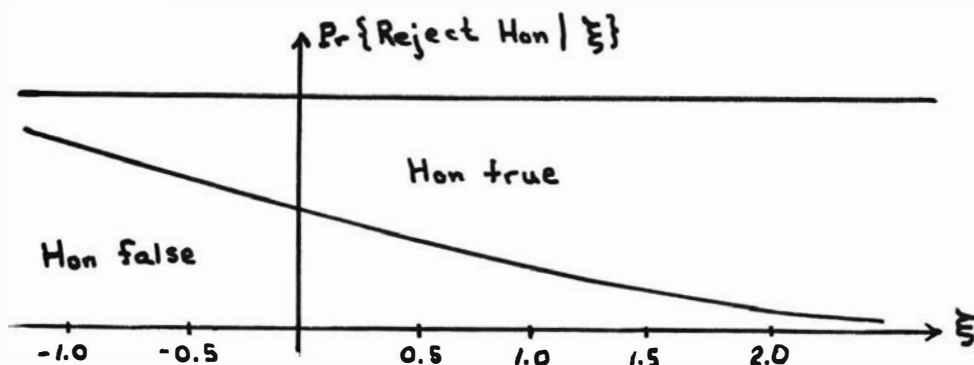


Fig. 17 Power of the current stopping test

Even when the slope is as small as -1.0 the stopping rule fails to diagnose a negative slope almost 25% of the time; still more damaging, it misrepresents a slope of + 1.0 as being negative equally often. In the example of Figure 16, the experimenter using the usual stopping rule runs a 25% risk of stopping at the first step, fully four units from the optimum. The case for an alternative is clear.

The near linearity of the power curve pins the blame on the variance of the statistic $(y_n - y_{n-1})$, and suggests that statistics which achieve a lower variance by incorporating more than two observations might improve the test procedure. Several possible alternatives using three, four, or more observations may be derived by fitting the model $\eta(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$ to the observations $y_n, y_{n-1}, \dots, y_{n-k}$ and combining the resulting expressions for $\hat{\gamma}_1, \hat{\gamma}_2$ to estimate the derivative at any chosen point $t = s_n$. For example, one might fit the model to the three observations y_n, y_{n-1}, y_{n-2} and derive an estimator which is unbiased for the slope at $s_n = n-1$. Rewriting the model

$$\begin{aligned}\eta(t) &= \gamma_0 + \gamma_1 [(t - s_n) + s_n] + \gamma_2 [(t - s_n) + s_n]^2 \\ &= [\gamma_0 + \gamma_1 s_n + \gamma_2 s_n^2] + [\gamma_1 + 2s_n \gamma_2](t - s_n) + \gamma_2 (t - s_n)^2 \\ &= \gamma_0^* + \gamma_1^*(t - s_n) + \gamma_2^*(t - s_n)^2,\end{aligned}$$

the derivative evaluated at $t = s_n$ reduces to $\gamma_1^* = \gamma_1 + 2s_n \gamma_2$

so that $\hat{\gamma}_1^*$ provides an unbiased estimator for the slope of the expected response function at $t = s_n$. For the particular case involving y_n, y_{n-1}, y_{n-2} with $s_n = n-1$, the derivation proceeds as follows:

$$X = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad \underline{y} = \begin{bmatrix} y_{n-2} \\ y_{n-1} \\ y_n \end{bmatrix}$$

$$\begin{bmatrix} \hat{\gamma}_0^* \\ \hat{\gamma}_1^* \\ \hat{\gamma}_2^* \end{bmatrix} = \begin{bmatrix} 3 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix}^{-1} \begin{bmatrix} y_n + y_{n-1} + y_{n-2} \\ y_n - y_{n-2} \\ y_n + y_{n-2} \end{bmatrix} = \begin{bmatrix} y_{n-1} \\ \frac{1}{2}(y_n - y_{n-2}) \\ (y_n - 2y_{n-1} + y_{n-2}) \end{bmatrix}$$

The estimator $\hat{\gamma}_1^* = \frac{1}{2}(y_n - y_{n-2})$ is unbiased for the slope $\gamma_1^* = \gamma_1 + 2(n-1)\gamma_2$ at $t=n-1$, and has a variance of only 0.5. The power curve for testing the null hypothesis of positive slope at $t=n-1$ is much more respectable when the test statistic $\frac{1}{2}(y_n - y_{n-2})$ is used, even at the implausible level $\alpha = .50$.

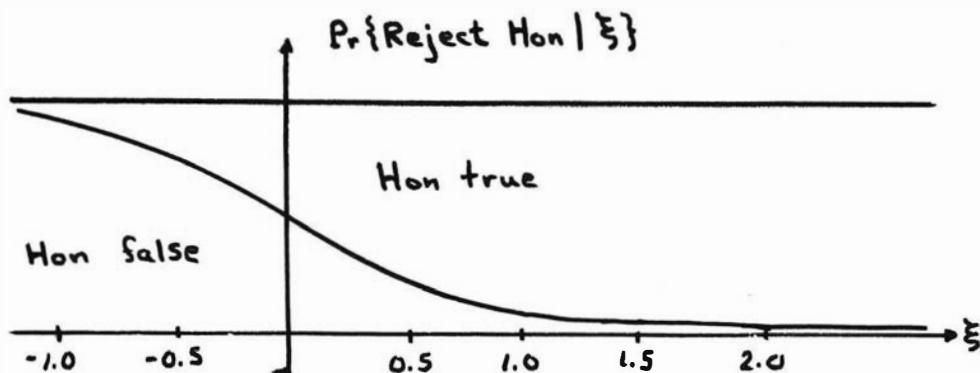


Fig. 18 Power of .5-level stopping test using $\frac{1}{2}(y_n - y_{n-2})$

One could obtain many other test statistics by applying the preceding method of derivation to various numbers of observations and different values of s_n (the point at which the test statistic is unbiased for the slope), but three factors limit the number of statistics which are likely to be practical. First, the experimenter cannot use a statistic which requires more than five or six observations for its computation, because he does not have that many observations available to him until late in the sequential decision phase.

Second, the use of several observations in the test procedure introduces undesirable correlations among the test statistics for successive tests. These correlations increase rapidly as the variances of the estimators decrease, so that even a procedure based on four points is seen to lose as much from the dependence among successive tests as it gains from the lower variances of the individual statistics. Third, the value of s_n cannot be much less than n without requiring the experimenter to take observations several steps ahead of the point at which he tests the slope. For example, if $s_n = n-2$, the experimenter must take observations at $t=n-1$ and $t=n$ before deciding whether the optimum stopping distance is $t=n-2$.

Table II summarizes the properties of the standard statistic $y_n - y_{n-1}$ and four alternatives which give promise of finding practical application.

Table II: Five Test Statistics

Statistic	No. Obs. Used	s_n	Variance	Formula
1	2	$n - \frac{1}{2}$	2.0	$y_n - y_{n-1}$
2	3	$n - 1$	0.5	$\frac{1}{2}(y_n - y_{n-2})$
3	4	$n - 1$	0.45	$\frac{1}{20}(11y_n - 7y_{n-1} - 3y_{n-2} - y_{n-3})$
4	4	$n - \frac{3}{2}$	0.2	$\frac{1}{10}(3y_n + y_{n-1} - y_{n-2} - 3y_{n-3})$
5	5	$n - 2$	0.1	$\frac{1}{10}(2y_n + y_{n-1} - y_{n-3} - 2y_{n-4})$

The dispersion matrices for five consecutive tests using each of the five procedures are given below.

$$\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

a) Dispersion matrix,
statistic 1

$$\begin{bmatrix} .5 & 0 & -.25 & 0 & 0 \\ 0 & .5 & 0 & -.25 & 0 \\ -.25 & 0 & .5 & 0 & -.25 \\ 0 & -.25 & 0 & .5 & 0 \\ 0 & 0 & -.25 & 0 & .5 \end{bmatrix}$$

b) Dispersion matrix,
statistic 2

$$\begin{bmatrix} .45 & -.0125 & -.185 & -.0275 & 0 \\ -.0125 & .45 & -.0125 & -.185 & -.0275 \\ -.185 & -.0125 & .45 & -.0125 & -.185 \\ -.0275 & -.185 & -.0125 & .45 & -.0125 \\ 0 & -.0275 & -.185 & -.0125 & .45 \end{bmatrix}$$

c) Dispersion matrix,
statistic 3

$$\begin{bmatrix} .2 & .05 & -.06 & -.09 & 0 \\ .05 & .2 & .05 & -.06 & -.09 \\ -.06 & .05 & .2 & .05 & -.06 \\ -.09 & -.06 & .05 & .2 & -.05 \\ 0 & -.09 & -.06 & .05 & .2 \end{bmatrix}$$

d) Dispersion matrix,
statistic 4

$$\begin{bmatrix} .1 & 0 & -.01 & 0 & .04 \\ 0 & .1 & 0 & -.01 & 0 \\ -.01 & 0 & .1 & 0 & -.01 \\ 0 & -.01 & 0 & .1 & 0 \\ .04 & 0 & -.01 & 0 & .1 \end{bmatrix}$$

e) Dispersion matrix, statistic 5

The dependence among tests detracts from the value of a single power curve as a gauge for evaluating a stopping rule whose application requires the use of several tests. Further, the power curve alone is of little help in fixing the optimum value of α , the level to be used in testing. The next section develops a criterion for measuring the performance of stopping rules based on each of the five test statistics, and for

III.C OPTIMALITY CRITERION FOR STOPPING RULES

The five test statistics of Table II generate five families of stopping rules indexed by α , the level of an individual test. For each model of the form $\eta(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$ the vector \underline{y} of single observations has a multivariate normal distribution:

$$\underline{y} \sim N(\underline{\eta}, I), \quad \text{where } \underline{\eta}' = (\eta(0), \eta(1), \eta(2), \dots).$$

Each test statistic \underline{z}_n is a linear compound of the observations, so a vector \underline{z} of successive test statistics can be written $C\underline{y}$.

$$\underline{z} = C\underline{y} \sim N(C\underline{\eta}, CC').$$

After each observation \underline{y}_n the experimenter tests H_{0n} :

"Slope at s_n positive," rejecting if and only if the value of the test statistic \underline{z}_n is less than some critical value \underline{z}_α determined by the choice of a test level α .

Associated with each model $\eta(t) = \gamma_0 + \gamma_1 t + \gamma_2 t^2$, each vector \underline{z} of test statistics, and each choice of α , one can define a discrete random variable s , the stopping distance. The random variable s ranges over the values s_n at which the experimenter tests the slope, with $p_n = \Pr\{s = s_n\}$ equal to the probability that the experimenter stops at s_n , i.e., that he first rejects the hypothesis of positive slope at $t = s_n$. The probability density for s can be computed from a family of multivariate normal integrals:

$$\begin{aligned} p_n &= \Pr\{s = s_n\} = \Pr\{\text{Accept } H_{01}, H_{02}, \dots, H_{0,n-1}, \text{Reject } H_{0n}\} \\ &= \Pr\{z_1 > z_\alpha, z_2 > z_\alpha, \dots, z_{n-1} > z_\alpha, z_n \leq z_\alpha\} \\ &= \int_{-\infty}^{z_\alpha} \int_{z_\alpha}^{\infty} \int_{z_\alpha}^{\infty} \cdots \int_{z_\alpha}^{\infty} f(z_1, z_2, \dots, z_{n-1}, z_n) dz_1 dz_2 \cdots dz_{n-1} dz_n, \end{aligned}$$

Where $f(z_1, z_2, \dots, z_{n-1}, z_n)$ is the multivariate normal density for an n -dimensional vector with mean $C\eta$ and variance CC' . The probability density function for the random variable s provides a complete and concise record of the performance of a stopping rule for particular values of $\gamma_0, \gamma_1, \gamma_2$ and α . For the example of Figure 18 ($\gamma_0 = 0, \gamma_1 = 1.0, \gamma_2 = -1$) an ideal stopping rule would concentrate the entire probability for s at the point of maximum expected response, i.e., $p_s = P\{s=5\} = 1.0$ so that $E(s) = t^*$, the ideal stopping point, and $\text{Var}(s) = 0$. As Figure 19 illustrates, the current stopping rule based on $y_n - y_{n-1}$ with $\alpha = .50$ disperses the probability over a wide and inappropriate range, so that $E(s) = 2.07$, $\text{Var}(S) = 1.88$.

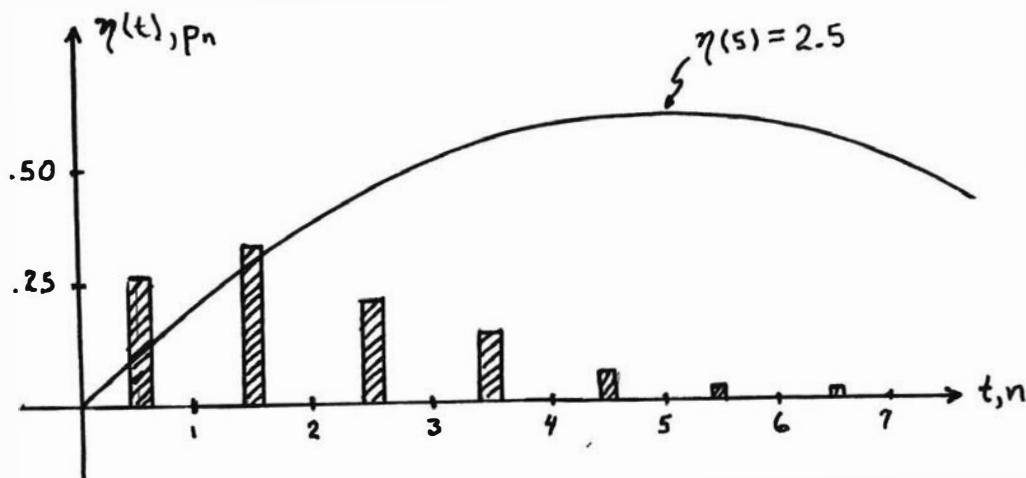


Fig. 19 Pdf and $\eta(t)$ for Rule 1, $\alpha = .5$, $\gamma_1 = 1.0$, $\gamma_2 = -1$

The simultaneous importance of variance and expectation suggests the use of a squared error loss function to define the optimum stopping rule.

$$\mathcal{L}(s | \alpha, \gamma_0, \gamma_1, \gamma_2) = E\{(s - t^*)^2\}$$

$$\begin{aligned}
&= E\{(b - E(s)) - [t^* - E(s)]\}^2\} \\
&= \text{Var}(s) + [E(s) - t^*]^2.
\end{aligned}$$

For given model parameters γ_1, γ_2 (γ_0 is obviously irrelevant) the optimum stopping rule would be that which minimizes the loss function $\mathcal{L}(s)$. The following conjecture facilitates the search for an optimum.

Conjecture: For a specified sequence of test statistics and fixed values of the model parameters γ_1, γ_2 , the squared error loss as a function of the single parameter α attains a unique minimum on the interval $[0, 1]$

The conjecture is a direct consequence of two lemmas, only one of which will be proved, concerning the variance $V(\omega) = \text{Var}(s | \alpha)$ and the squared bias $B(\omega) = [E(s | \alpha) - t^*]^2$. The variance is conjectured to be a monotone decreasing function of α , and the squared bias is shown to be monotone increasing on either side of a minimum of zero at a unique α_u so that the sum $\mathcal{L}(\omega) = V(\omega) + B(\omega)$ must possess a unique minimum at some $\alpha^* \geq \alpha_u$ (Figure 20).

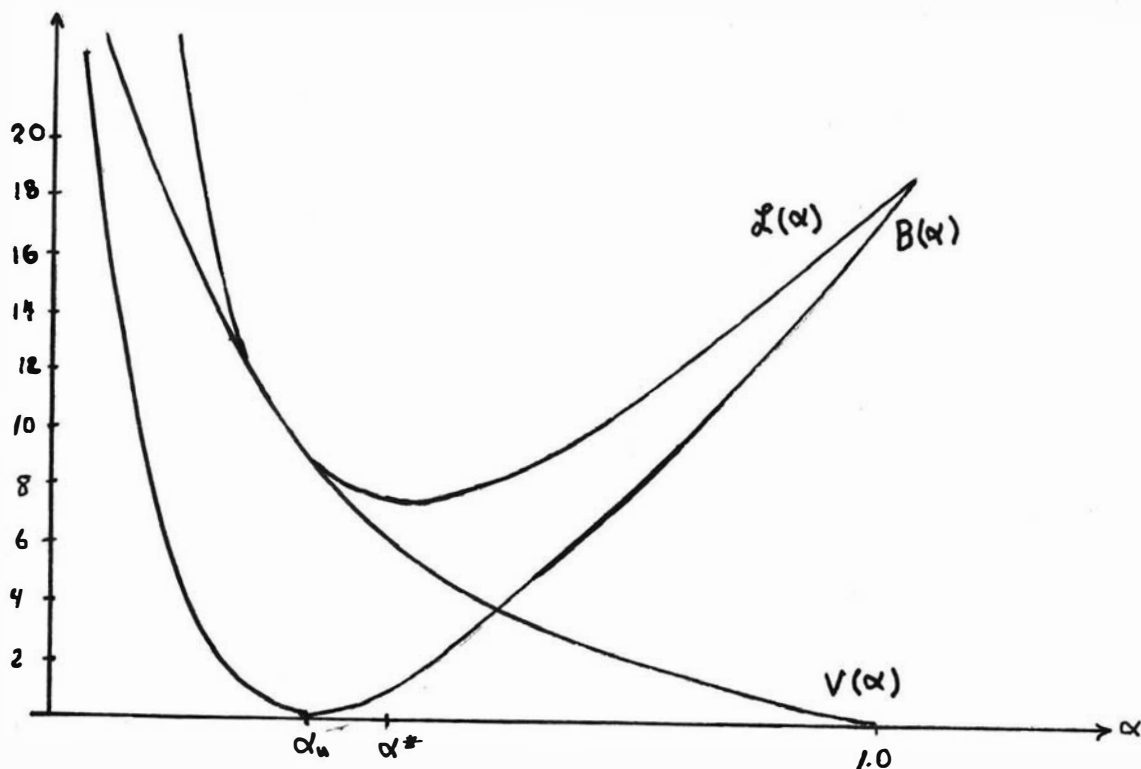


Fig. 20 $L(\alpha)$ for the standard stopping rule

$$\gamma_1 = 1.0, \gamma_2 = -.1, t^* = 5$$

The following notation and immediate results are needed to establish the result concerning the bias. As before, let s_n denote the value of t at which z_n is unbiased for the slope $\partial \eta(t) / \partial t$, i.e.,

$$E(z_n) = \left. \frac{\partial \eta(t)}{\partial t} \right|_{t=s_n} = \gamma_1 + 2s_n \gamma_2.$$

In particular, if $z_n = y_n - y_{n-1}$, $s_n = n - .5$; if $z_n = \frac{1}{2}(y_n - y_{n-2})$ or $z_n = \frac{1}{2}(11y_n - 7y_{n-1} - 3y_{n-2} - y_{n-3})$, $s_n = n - 1$. Then

$$E(s) = \sum_{n=1}^{\infty} p_n s_n$$

85

$$\text{Var}(s) = \sum_{n=1}^{\infty} p_n s_n^2 - [E(s)]^2 \quad (\text{III.C.1})$$

where, as before,

$$p_n = \int_{-\infty}^{z_n} \int_{z_n}^{\infty} \int_{z_n}^{\infty} \cdots \int_{z_n}^{\infty} f(z_1, z_2, \dots, z_{n-1}, z_n) dz_1 dz_2 \cdots dz_{n-1} dz_n.$$

Lemma 1: $B(\alpha)$ attains a unique minimum at some point $\alpha_u \in [0, 1]$

and is a strictly increasing function of the distance $|\alpha - \alpha_u|$

on either side of α_u . If $t^* > s_1$, $B(\alpha_u) = 0$ and $\alpha_u \in (0, 1)$;

if $t^* \leq s_1$, $B(\alpha_u) = (s_1 - t^*)^2$ and $\alpha_u = 1$.

Proof: Let $b(\alpha) = E(s|\alpha) - t^*$, so that $B(\alpha) = [b(\alpha)]^2$.

It suffices to show that $b(\alpha)$ is a strictly decreasing

function of α , with $b(0) > 0$, and $b(1) < 0$ if and

only if $t^* > s_1$. From III.C.1,

$$b(\alpha) = E(s|\alpha) - t^* = \sum_{n=1}^{\infty} p_n s_n - t^*.$$

Let $P_n = P_r\{s \geq s_n\} = 1 - \sum_{j=1}^{n-1} P_j$; let $s_0 = 0$. Then

$$b(\alpha) = \sum_{n=1}^{\infty} P_n (s_n - s_{n-1}) - t^*.$$

But $P_n = \int_{z_n}^{\infty} \cdots \int_{z_n}^{\infty} f(z_1, \dots, z_{n-1}) dz_1 \cdots dz_{n-1}$. As α

increases, z_n increases, hence P_n decreases, and $b(\alpha)$ decreases.

Further, as $\alpha \rightarrow 0$, $z_n \rightarrow -\infty$, hence $P_n \rightarrow 1$ for every n , and $b(\alpha) \rightarrow \infty$.

Finally, for $\alpha = 1$, $P_1 = 1$ and $P_n = 0$ for $n > 1$, so $b(\alpha) = s_1 - t^*$,

which is negative if and only if $s_1 < t^*$.

Lemma 2: $V(\alpha)$ is a strictly decreasing function of α , with $V(1)=0$ and $\lim_{\alpha \rightarrow 0} V(\alpha) = \infty$.

This result appears without proof, although an intuitive feel for the truth of the lemma can be gleaned from an analogy with the negative binomial distribution. Each test of the sequential decision phase corresponds to a single trial of a negative binomial experiment, with acceptance considered a success and rejection considered a failure. The stopping distance s then corresponds to the negative binomial random variable whose value is the number of successes before the first failure. It is easily established that the variance of a negative binomial random variable increases to infinity as the probability of failure decreases to zero. By analogy, one would expect the variance of the stopping distance to increase to infinity as the test level α , and hence the probability of rejecting any given hypothesis, decreases to zero.

The argument by analogy breaks down because the probability of rejecting the hypothesis H_{0n} : "Slope at s_n positive" is a function not only of α but also of n . Nevertheless, the lemma was seen to be true in all of the more than 600 specific examples considered.

III.D APPLICATION OF THE OPTIMALITY CRITERION

Two difficulties hinder the application of the loss function in determining an optimum stopping rule. First and

most immediate, obtaining numerical values of the loss function for particular stopping rules requires the evaluation of a large class of multivariate normal integrals. The loss is best obtained as the sum of the variance V and the squared bias B , which are computed from the formulas of equations III.C.i as weighted infinite sums. The weights P_n are n -variate integrals which must be evaluated numerically, either by quadrature methods or by Monte Carlo. Appendix II on the Multivariate Integral Problem discusses the relative merits of several evaluation schemes. In the end the integrals were evaluated using straightforward random sampling on an IBM 1130; the chosen sample size of 10,000 severely restricted the number of cases considered, although it was possible to evaluate several integrals simultaneously from a single sample. The second and more uncompromising problem resulted from the dependence of the loss function on the model parameters γ_1, γ_2 . Although the conjecture of III.C insures the existence for fixed γ_1, γ_2 of a unique optimum test level α^* for any sequence of test statistics, the optimum level α^* and even the preferred test statistic must be expected to vary with the parameters. Consequently it is impossible to avoid evaluating the loss function associated with each test statistic at several α -levels for a large number of parameter combinations.

Taking the five test statistics of Table II as a basis, five families of stopping rules were defined, with each to

be evaluated using the loss function at several levels of α for values of γ_1, γ_2 ranging over a grid representing models considered typical of response surface problems.

Rule 1

$$\begin{aligned} z_1 &= y_1 - y_0 & s_1 &= \frac{1}{2} \\ z_2 &= y_2 - y_1 & s_2 &= 1\frac{1}{2} \\ &\vdots & & \\ z_n &= y_n - y_{n-1} & s_n &= n - \frac{1}{2} \end{aligned}$$

$$C_1 = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}$$

Rule 2

$$\begin{aligned} z_1 &= y_1 - y_0 & s_1 &= \frac{1}{2} \\ z_2 &= \frac{1}{2}(y_2 - y_0) & s_2 &= 1 \\ &\vdots & & \\ z_n &= \frac{1}{2}(y_n - y_{n-2}) & s_n &= n - 1 \end{aligned}$$

$$C_2 = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 & 0 \\ -.5 & 0 & .5 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -.5 & 0 & .5 \end{bmatrix}$$

Rule 3

$$\begin{aligned} z_1 &= y_1 - y_0 & s_1 &= \frac{1}{2} \\ z_2 &= \frac{1}{2}(y_2 - y_0) & s_2 &= 1 \\ z_3 &= \frac{1}{20}(11y_3 - 3y_2 - 7y_1 - y_0) & s_3 &= 2 \\ &\vdots & & \\ z_n &= \frac{1}{20}(11y_n - 3y_{n-1} - 7y_{n-2} - y_{n-3}) & s_n &= n - 1 \end{aligned}$$

$$C_3 = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -.5 & 0 & .5 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -.05 & -.35 & -.15 & .55 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -.05 & -.35 & -.15 & .55 \end{bmatrix}$$

Rule 4

$$z_1 = y_1 - y_0$$

$$s_1 = \frac{1}{2}$$

$$z_2 = \frac{1}{2}(y_2 - y_0)$$

$$s_2 = 1$$

$$z_3 = \frac{1}{10}(3y_3 + y_2 - y_1 - 3y_0)$$

$$s_3 = 1\frac{1}{2}$$

$$\vdots$$

$$\vdots$$

$$z_n = \frac{1}{10}(3y_n + y_{n-1} - y_{n-2} - 3y_{n-3})$$

$$s_n = n - 1\frac{1}{2}$$

$$C_4 = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -.5 & 0 & .5 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -.3 & -.1 & .1 & .3 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -.3 & -.1 & .1 & .3 \end{bmatrix}$$

Rule 5

$$z_1 = y_1 - y_0$$

$$s_1 = \frac{1}{2}$$

$$z_2 = \frac{1}{2}(y_2 - y_0)$$

$$s_2 = 1$$

$$z_3 = \frac{1}{10}(3y_3 + y_2 - y_1 - 3y_0)$$

$$s_3 = 1\frac{1}{2}$$

$$z_4 = \frac{1}{10}(2y_4 + y_3 - y_1 - 2y_0)$$

$$s_4 = 2$$

$$\vdots$$

$$\vdots$$

$$z_n = \frac{1}{10}(2y_n + y_{n-1} - y_{n-3} - 2y_{n-4})$$

$$s_n = n - 2$$

$$C_5 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ -.5 & 0 & .5 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ -.3 & -.1 & .1 & .3 & 0 & \cdots & 0 & 0 & 0 & 0 & 0 \\ -.2 & -.1 & 0 & .1 & .2 & \cdots & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -.2 & -.1 & 0 & .1 & .2 \end{bmatrix}$$

Originally 121 parameter pairs (γ_1, γ_2) were chosen to cover a grid with γ_1 ranging from 0.0 to 2.0 and γ_2 ranging from -0.01 to -2.0. All but 48 of these cases were then eliminated as atypical because they generated values of t^* which were either less than 1.0 or greater than 10.0 (Table III).

Table III: Values of t^* for various											
$\gamma_2 \backslash \gamma_1$	0.	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
-0.01	0.	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100.0
-0.02	0.	5.0	10.0	15.0	20.0	25.0	30.0	35.0	40.0	45.0	50.0
-0.03	0.	3.3	6.7	10.0	13.3	16.7	20.0	23.3	26.7	30.0	33.3
-0.05	0.	2.0	4.0	6.0	8.0	10.0	12.0	14.0	16.0	18.0	20.0
-0.10	0.	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
-0.15	0.	0.7	1.3	2.0	2.7	3.3	4.0	4.7	5.3	6.0	6.7
-0.25	0.	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
-0.50	0.	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
-0.75	0.	.13	.27	.40	.53	.67	.80	.93	1.07	1.2	1.33
-1.00	0.	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
-2.00	0.	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50

For each test statistic and parameter pair, the loss was evaluated at six levels of α and approximated by a polynomial whose minimum determined the optimum level α^* . The losses for $\alpha = .1, .3, .5, .7, .9$ were computed by numerical integration; the loss for $\alpha = 1.0$ is given by $(s - t^*)^2$, since $\alpha = 1$ forces the experimenter to reject the first hypothesis.

Chapter IV presents the optimum levels of α and the associated losses, and attempts to assimilate from the diverse values some general conclusions of practical significance.

CHAPTER IV: RESULTS AND CONCLUSIONS

IV.A NUMERICAL RESULTS

The results of this chapter lead to a modification of the standard stopping rule which breathes new life into the method of steepest ascent, but which is unfortunately not sufficient to effect a total resuscitation. As expected, there is no uniformly optimum stopping rule, nor even a uniformly best test statistic; both the optimum statistic and the optimum test level depend on the model parameters γ_1 and γ_2 . One can, however, adopt a compromise rule which works reasonably well for all parameter combinations by testing at the 25% level with the statistic $z_n = \frac{1}{2}(y_n - y_{n-2})$. In many cases the loss incurred by the compromise rule is less than half that incurred by the existing rule; in the few cases where the compromise rule is less efficient, the difference in losses is small. Nevertheless, the compromise rule is nearly always inferior to the parameter-dependent optimum rule, and the optimum itself is not always desirable.

The chapter ends with a frankly subjective and imprecise comparison of the relative merits of steepest ascent using the compromise rule and pseudo-steepest ascent using the simplex method. One can easily imagine situations in which one or the other would be preferable, suggesting that the two methods are complimentary rather than competitive,

and that the best region seeking procedure might offer the experimenter a combination of the two.

The numerical results occupy tables IV and V and Figures 21-25. In order to spare the reader the unnecessary burden of sifting through an intimidating array of numbers in search of meaningful generalities, the relevant information which can be extracted from the tables and graphs has been summarized in what follows. Conclusions, practical recommendations, and discussion will all depend on the summaries rather than on the tabulated results, so only the skeptic, the masochist, or the graduate committee member need plunge into the morass of raw data.

Table IV compares the minimum losses which are obtained when Rules 1-5 are used at their optimum levels α^* . The corresponding losses for the standard rule, i.e., Rule 1 with $\alpha = .50$, occupy the extreme right-hand column. Rows of the table are indexed by t^* ; where more than one parameter pair (γ_1, γ_2) yields the same value for t^* , losses corresponding to lower values of γ_1 and γ_2 appear first. The results support the following generalizations.

1. Relative to the optimum, the standard rule performs very badly, incurring losses as high as 76.31, with one third of the cases considered resulting in losses above 10.0. The losses are particularly high for $t^* \geq 5.0$; for low t^* , between 1.0 and 2.0, the losses are also low, less than 2.0.

2. The minimum loss for each of the rules is roughly an increasing function of t^* . For $t^* < 2.0$, the losses are all below 1.0; as t^* increases from 2.0 to 4.0 the losses increase from .72 to 4.98; as t^* increases to 10.0 the losses increase to values around 20.0. When several parameter pairs (γ_1, γ_2) yield equal values of t^* , the loss generally increases with the parameters.

3. For fixed γ_1 and γ_2 , the minimum losses are remarkably uniform from rule to rule. One might have predicted Rule 5, with the lower variance of its test statistic, to perform significantly better than the other rules; likewise, one might have predicted Rule 1, with its statistic having a variance twenty times greater than that of Rule 5, to perform significantly worse than the others. To an extent the results support both predictions when $t^* \leq 5$, but the overriding impression is of uniformity.

4. The results are generally uninteresting for $t^* < 3.0$. The losses are all low, and no single rule consistently performs better or worse than the others.

Figures 21-25, which plot the optimum test level α^* for each of the five rules as a function of t^* for the 48 cases considered, reveal the following results.

5. The dependence of the optimum test level α^* on the parameters γ_1 and γ_2 can be reexpressed as a dependence on the single parameter $t^* = -\gamma_1 / (2\gamma_2)$ provided $\gamma_2 \geq .10$. In each of the figures, points corresponding to cases where $\gamma_2 < .10$ are plotted as x's to distinguish them as exceptions.

Table IV: Minimum Losses for Rules 1-5

t*	γ_1	γ_2	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Standard
1.0	.20	-.10	.21	.25	.24	.24	.23	.72
	1.00	-.50	.24	.24	.23	.21	.20	.56
	2.00	-1.00	.24	.23	.23	.15	.15	.54
1.1	1.60	-.75	.27	.29	.28	.22	.20	.54
1.2	.60	-.25	.37	.43	.39	.43	.39	.60
	1.20	-.50	.36	.39	.42	.38	.30	.58
	1.80	-.75	.32	.38	.41	.30	.24	.54
1.3	.40	-.15	.50	.57	.57	.58	.58	.69
	2.00	-.75	.37	.46	.47	.41	.26	.54
1.4	1.40	-.50	.47	.54	.55	.55	.40	.61
1.6	.80	-.25	.70	.80	.80	.84	.83	.73
	1.60	-.50	.58	.63	.66	.70	.53	.65
1.8	1.80	-.50	.68	.73	.73	.71	.70	.72
2.0	.20	-.05	1.16	1.46	1.51	1.64	1.74	1.30
	.40	-.10	1.23	1.44	1.45	1.50	1.57	1.26
	.60	-.15	1.12	1.29	1.31	1.42	1.41	1.14
	1.00	-.25	.99	1.13	1.13	1.17	1.13	1.00
	2.00	-.50	.75	.80	.81	.72	.77	.76
2.4	1.20	-.25	1.30	1.41	1.46	1.46	1.35	1.36
2.7	.80	-.15	1.80	2.00	2.08	2.21	2.12	2.09
2.8	1.40	-.25	1.58	1.60	1.67	1.67	1.46	1.81
3.0	.60	-.10	2.68	2.86	2.81	3.01	3.01	3.22
3.2	1.60	-.25	1.88	1.72	1.85	1.80	1.59	2.31
3.3	.20	-.03	3.51	3.86	3.96	4.32	4.68	4.98
	1.00	-.15	2.71	2.67	2.83	2.88	2.66	3.50
3.6	1.80	-.25	2.18	1.89	1.95	1.79	1.60	2.86
4.0	.40	-.05	5.27	5.25	5.51	5.79	5.93	7.52
	.80	-.10	4.55	4.31	4.30	4.35	4.25	6.29
	1.20	-.15	3.65	3.25	3.45	3.25	3.04	5.26

Table IV, (cont'd)

t*	γ_1	γ_2	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Standard
4.0	2.00	-.25	2.36	1.94	1.90	1.79	1.55	3.37
4.7	1.40	-.15	4.39	3.60	3.58	3.55	3.29	7.17
5.0	.20	-.02	9.90	10.02	9.79	9.96	10.18	14.48
	1.00	-.10	6.70	5.66	5.69	5.46	5.05	10.46
5.3	1.60	-.15	5.23	3.85	3.90	3.77	3.33	9.14
6.0	.60	-.05	12.08	10.54	10.86	9.96	9.81	20.06
	1.20	-.10	8.86	6.53	6.71	6.06	5.47	15.49
	1.80	-.15	6.07	4.01	3.89	3.78	3.40	11.27
6.7	.40	-.03	17.07	15.63	16.33	14.34	13.79	28.06
	2.00	-.15	6.71	4.09	3.81	3.72	3.38	13.41
7.0	1.40	-.10	10.92	7.35	7.33	6.23	6.20	21.09
8.0	.80	-.05	18.88	12.15	14.89	12.96	12.09	38.47
	1.60	-.10	12.79	8.09	7.15	6.79	6.24	27.09
9.0	1.80	-.10	14.26	8.28	6.81	6.93	6.52	32.80
10.0	.20	-.01	14.51	13.40	15.88	16.58	19.53	76.31
	.40	-.02	16.62	18.56	18.80	20.91	20.30	73.48
	.60	-.03	21.46	19.61	19.88	18.45	16.57	70.44
	1.00	-.05	21.19	16.08	15.01	13.47	20.14	61.75
	2.00	-.10	14.77	8.50	7.41	7.24	6.87	38.58

IV.B PRACTICAL RECOMMENDATIONS

The preceding body of results suggests three recommendations to the experimenter. The first is a compromise stopping rule for use when the experimenter has a reliable estimate of the error variance σ^2 . The second proposes an alternative which, because it costs the experimenter more than the first compromise rule, is recommended only for use with four or more independent variables where the large cost of first-order designs requires that the experimenter obtain maximum use from each, even at the expense of one or two additional observations. The third recommendation extends the previous two to cover the case of an experiment in which σ^2 is unknown and must be estimated from the observations. Of the three proposals, only the first two find solid support in the available data; the third, though suggested by the data, is nevertheless more speculative.

The table and particularly the graphs of IV.A confront the experimenter with an apparent dilemma: regardless of the test statistic he uses, he cannot find a test level α which works equally well for all values of t^* . If he chooses to operate with α at a high level to insure early stopping and minimum loss when t^* is small, he risks large losses due to premature stopping if t^* should turn out to be large; conversely, if he chooses a relatively low level for α to permit the taking of several observations when t^* is large, he may overshoot the optimum stopping point and incur a loss

which is triple or even quadruple the minimum if t^* should turn out to be small. The compromise rule which partially resolves this dilemma takes its justification from several facts. First, of the two evils described, the lesser by far results from choosing a relatively low value for α , for even if losses for small t^* are quadruple the minimum, they remain below 3.0, so that the experimenter can still expect to stop his single observations in the general vicinity of the optimum stopping point. Second, the experimenter need not be concerned with the extreme cases $t^* = 10.0$ and $t^* = 1.0$. He is rarely lucky enough to take as many as ten single observations before reaching the optimum stopping point, and rarely unlucky enough to need a new first-order design after only one or two observations. In the typical application he expects to take between four and eight single observations before refitting, with six as a very rough average; this expectation would lead him intuitively to select a test level of .25, which is nearly optimum for t^* between 5.0 and 7.0. Third, $\alpha = .25$ corresponds to the flattest segments of the graphs in Figures 21-25, indicating that no other value of α would work as well over as large a range of t^* .

Table V compares the performances of the five rules at the 25% level with that of the standard rule and the parameter-dependent optimum:

6. As in the case of the minimum losses, there is remarkably little variation in the results from rule to rule. The optimum level α^* appears to be determined by t^* independent of the class of test statistics used.

7. Each of the five figures reveals the intuitive result that as the optimum stopping distance increases, the most effective test level decreases. For $1 \leq t^* < 4$, the decrease is rapid; for $4 \leq t^* \leq 10$, the decrease is gradual, and can be expressed as a linear function of t^* .

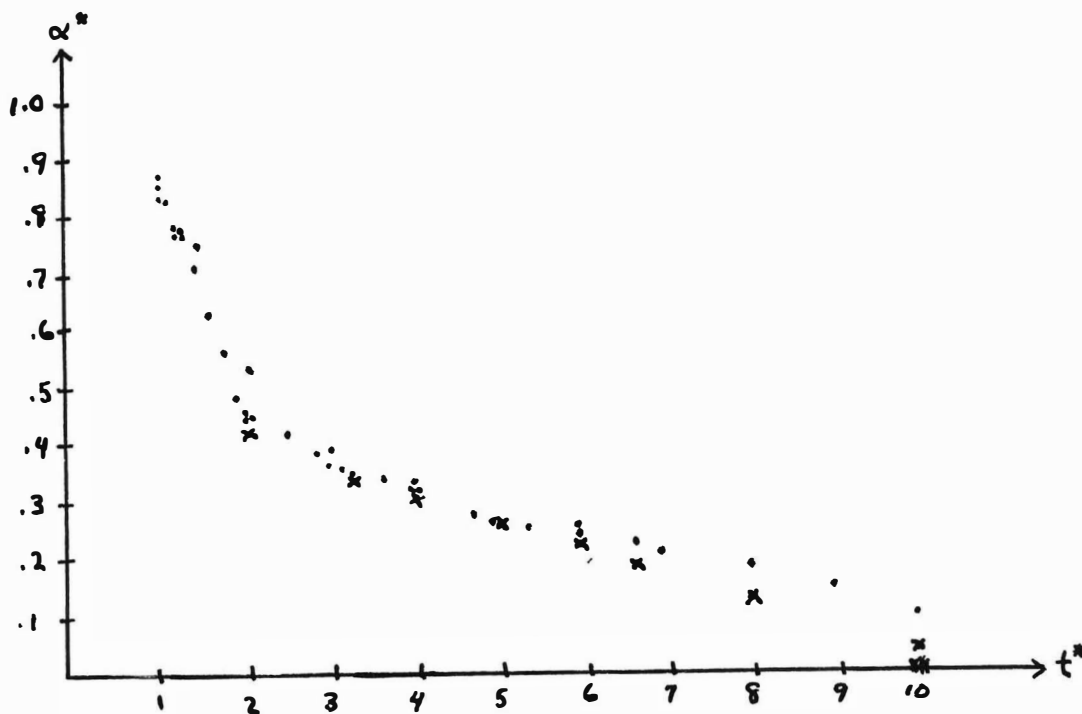
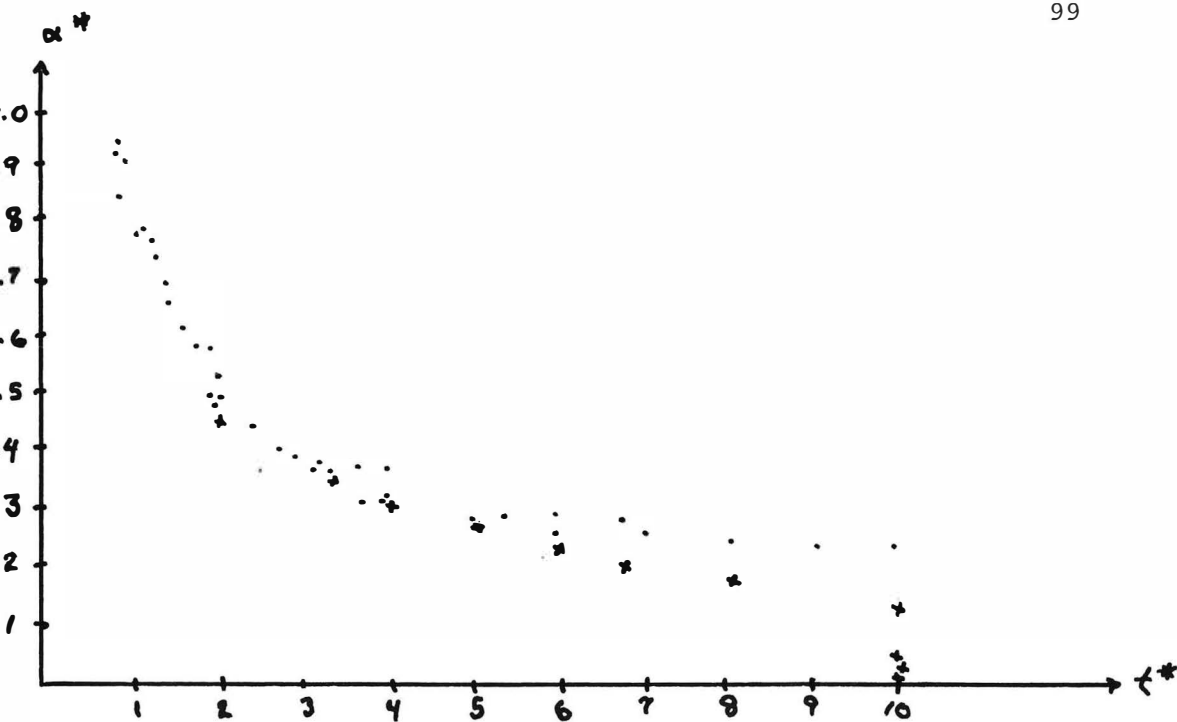
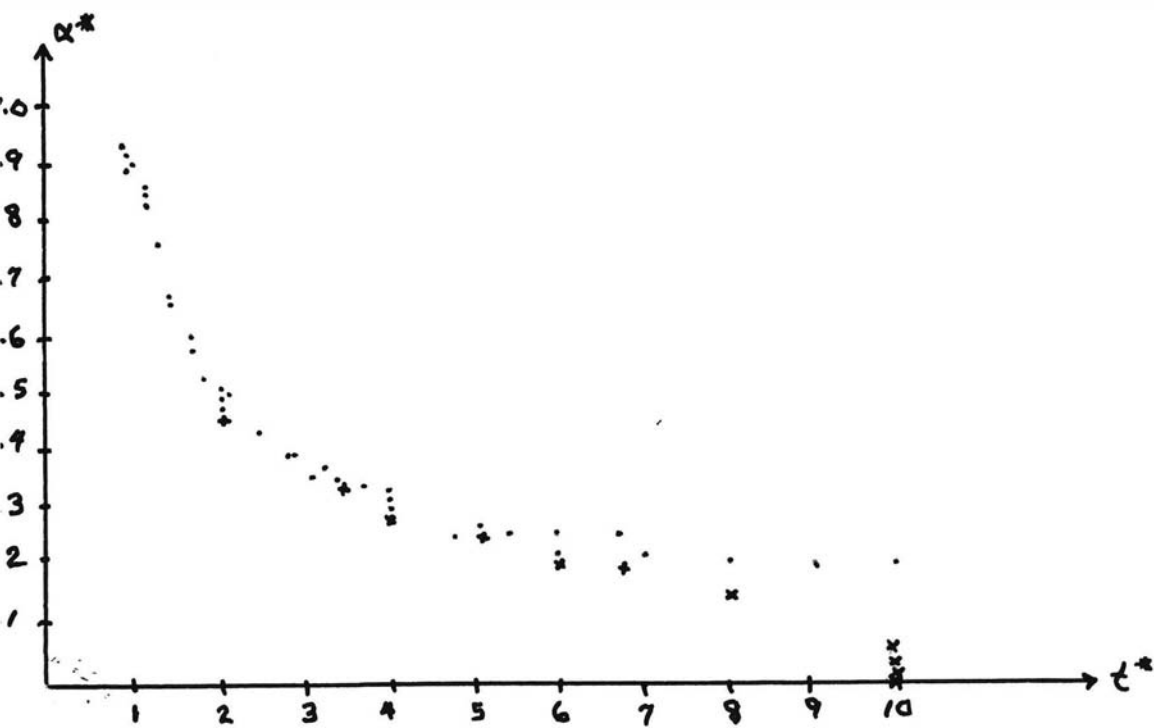
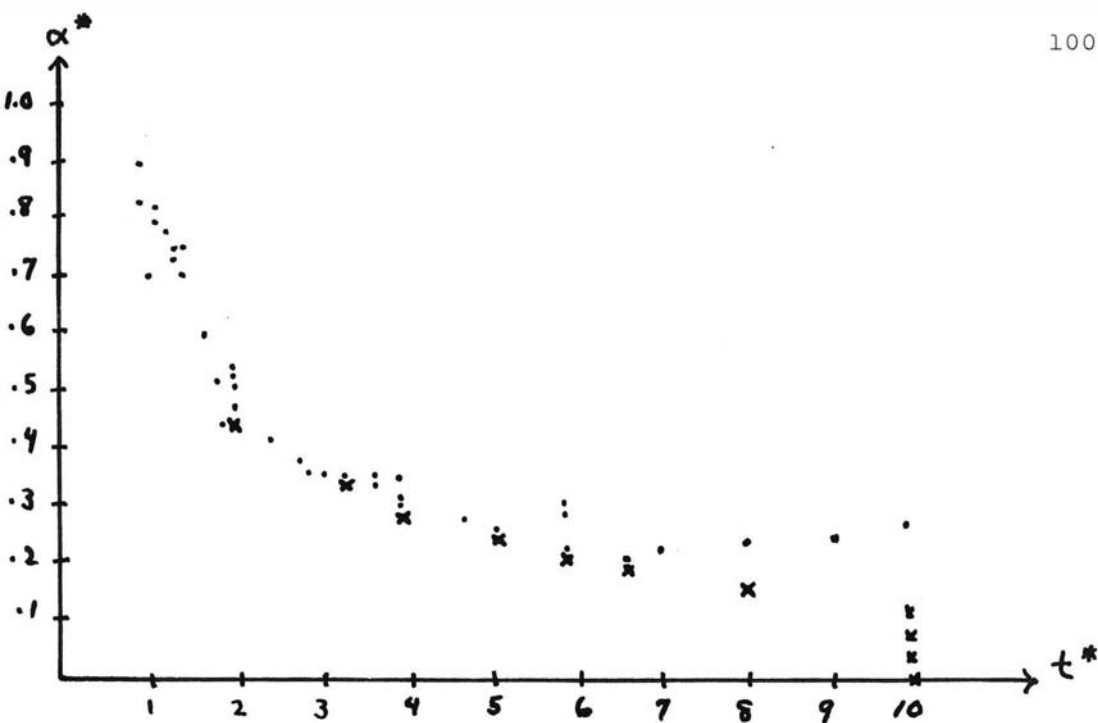
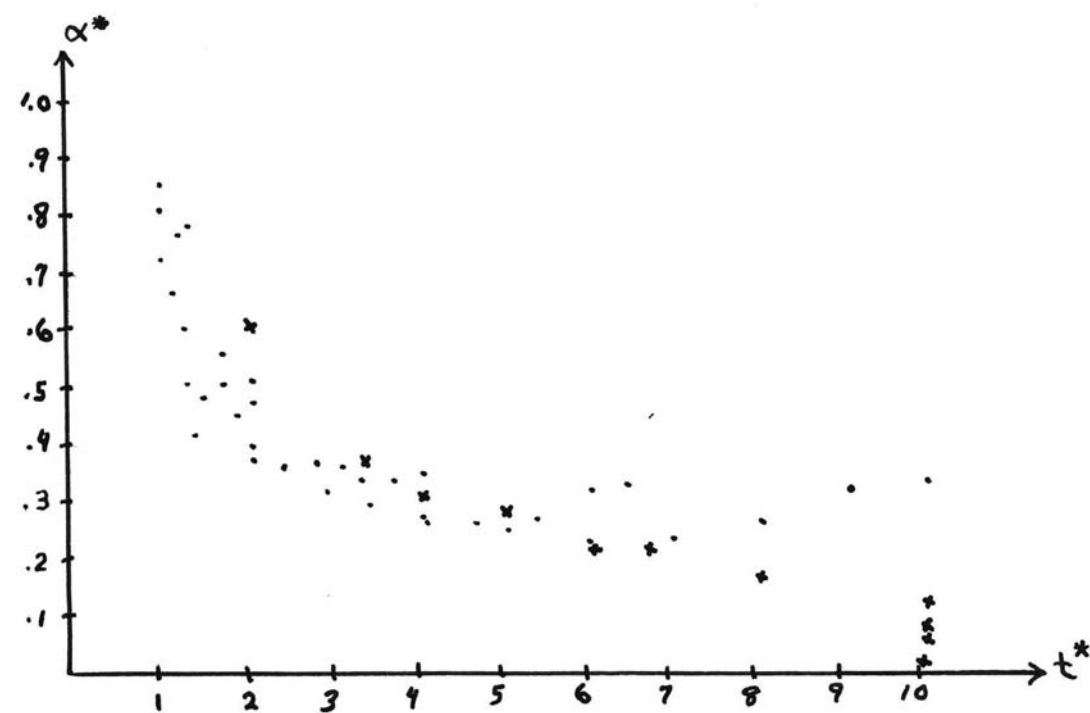


Fig. 21 Optimum level α^* for Rule 1

Fig. 22 Optimum level α^* for Rule 2Fig. 23 Optimum level α^* for Rule 3

Fig. 24 Optimum level α^* for Rule 4Fig. 25 Optimum level α^* for Rule 5

8. At the 25% level, Rule 1 is clearly inferior to the other four rules, especially for $t^* \geq 5.0$

9. The losses associated with Rules 2 and 3 are almost identical for all values of t^* . Since the test statistic for Rule 2 is simpler and involves fewer observations than that for Rule 3, Rule 2 is preferable.

10. Rules 4 and 5 both require the experimenter to take two observations ahead of the point where he tests the slope. With the exception of three cases, Rule 5 always incurs smaller losses than Rule 4, and should be preferred when the experimenter is willing to invest in the "extra" observations beyond the point under test.

The results of Table V suggest that the experimenter should use either Rule 2 or Rule 5, with the choice between them depending on his willingness to invest in the additional observation required by Rule 5. Both of the following recommendations lead to a large improvement in the standard stopping rule.

Recommendation 1: If the experimenter knows σ^2 from some previous reliable estimate, he should use Rule 2 at the 25% level to decide when to stop his single observations:

$$n = 1 \quad \text{Reject } H_{01} \quad y_1 - y_0 < -1.96\sigma$$

$$n \geq 2 \quad \text{Reject } H_{0n} \quad \frac{1}{2}(y_n - y_{n-1}) < -.48\sigma$$

Recommendation 2: If the number of independent variables in an experiment is large, so that the cost of a first-order design relative to the cost of the single observations of

of the decision phase is also large, the experimenter should use Rule 5 at the 25% level. The need to extract maximum efficiency from the costly first-order design justifies the expense of the extra observation required by Rule 5.

$$n=1 \quad \text{Reject } H_{01} \quad y_1 - y_0 < -.96\sigma$$

$$n=2 \quad \text{Reject } H_{02} \quad \frac{1}{2}(y_2 - y_0) < -.48\sigma$$

$$n=3 \quad \text{Reject } H_{03} \quad \frac{1}{10}(3y_3 + y_2 - y_1 - 3y_0) < -.30\sigma$$

$$n \geq 4 \quad \text{Reject } H_{0n} \quad \frac{1}{10}(2y_n + y_{n-1} - y_{n-3} - 2y_{n-4}) < -.22\sigma$$

Table V: Loss for $\alpha = .25$

t*	γ_1	γ_2	Stand- ard	Mini- mum	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
1.0	.20	-.10	.72	.21	4.45	3.25	3.14	2.68	2.29
	1.00	-.50	.56	.20	1.62	1.09	1.06	.79	.44
	2.00	-1.00	.54	.15	1.09	.75	.75	.31	.23
1.1	1.60	-.75	.54	.20	1.27	.81	.82	.55	.31
1.2	.60	-.25	.60	.37	2.36	1.67	1.67	1.39	1.03
	1.20	-.50	.58	.30	1.68	1.10	1.08	.87	.43
	1.80	-.75	.54	.24	1.26	.84	.82	.67	.31
1.3	.40	-.15	.69	.50	3.05	2.33	2.37	2.02	1.64
	2.00	-.75	.54	.26	1.27	.87	.84	.76	.30
1.4	1.40	-.50	.61	.40	1.70	1.16	1.12	.90	.53
1.6	.80	-.25	.73	.70	2.29	1.87	1.80	1.47	1.18
	1.60	-.50	.65	.53	1.67	1.18	1.19	.89	.71
1.8	1.80	-.50	.72	.68	1.63	1.27	1.23	.91	.85
2.0	.20	-.05	1.30	1.16	4.52	3.80	3.97	4.12	3.75
	.40	-.10	1.26	1.23	3.85	3.11	2.96	2.89	2.45
	.60	-.15	1.14	1.12	2.91	2.46	2.45	2.24	1.88
	1.00	-.25	1.00	.99	2.20	1.92	1.87	1.64	1.29
	2.00	-.50	.76	.72	1.63	1.28	1.22	.96	.83
2.4	1.20	-.25	1.36	1.30	2.43	2.00	2.05	1.80	1.46
2.7	.80	-.15	2.09	1.80	3.06	2.78	2.78	2.65	2.32
2.8	1.40	-.25	1.81	1.46	2.54	2.14	2.17	1.87	1.52
3.0	.60	-.10	3.22	1.68	4.24	3.65	3.56	3.56	3.29
3.2	1.60	-.125	2.31	1.59	2.66	2.16	2.22	2.03	1.68
3.3	.20	-.03	4.98	3.51	5.43	5.41	5.50	5.84	5.94
	1.00	-.15	3.50	2.66	3.56	3.12	3.27	3.12	2.70
3.6	1.80	-.25	2.86	1.60	2.79	2.46	2.28	2.11	1.73
4.0	.40	-.05	7.52	5.25	6.14	5.66	6.03	6.21	6.17
	.80	-.10	6.29	4.25	5.36	4.64	4.50	4.50	4.28
	1.20	-.15	5.26	3.04	4.14	3.51	3.71	3.39	3.06

Table V (cont'd)

t*	γ_1	γ_2	Stand- ard	Mini- mum	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
	2.00	-.25	3.37	1.55	2.87	2.39	2.22	2.16	1.75
4.7	1.40	-.15	7.17	3.29	4.63	3.78	3.58	3.66	3.30
5.0	.20	-.02	14.48	9.79	10.12	10.21	10.06	10.26	10.41
	1.00	-.10	10.46	5.05	6.94	5.74	5.72	5.47	5.05
5.3	1.60	-.15	9.14	3.33	5.31	3.99	3.99	3.91	3.35
6.0	.60	-.05	20.06	9.81	12.12	10.68	10.96	10.08	9.95
	1.20	-.10	15.49	5.47	8.86	6.53	6.73	6.07	5.59
	1.80	-.15	11.27	3.40	6.09	4.17	3.94	3.98	3.58
6.7	.40	-.03	28.06	13.79	17.77	16.35	16.75	14.80	14.22
	2.00	-.15	13.41	3.38	6.71	4.21	3.86	3.96	3.87
7.0	1.40	-.10	21.09	6.20	11.03	7.36	7.40	6.25	6.23
8.0	.80	-.05	38.47	12.09	21.80	15.52	16.88	14.10	13.37
	1.60	-.10	27.09	6.24	13.39	8.12	7.29	6.79	6.24
9.0	1.80	-.10	32.80	6.52	15.60	8.32	7.10	6.95	6.80
10.0	.20	-.01	76.31	13.40	54.51	51.99	50.81	47.45	44.34
	.40	-.02	73.48	16.62	49.12	43.45	43.01	36.55	32.99
	.60	-.03	70.44	16.57	43.90	35.78	34.95	28.55	25.29
	1.00	-.05	61.75	15.01	33.64	21.99	23.07	17.17	15.51
	2.00	-.10	38.58	6.87	17.70	8.59	7.51	7.47	7.65

Table VI summarizes the numerical argument for adoption of the two recommendations. For $t^* \geq 4$, the percent by which each of the new rules reduces the loss from that of the old rule toward the minimum is computed:

$$\% \text{ Improvement} = \left\{ \frac{\text{old loss} - \text{new loss}}{\text{old loss} - \text{minimum}} \right\} \times 100$$

Measured in this way, the improvement resulting from either proposed modification of the stopping rule is generally near 90%, often higher than 95% for Rule 5. The exceptional cases in which the percent improvement falls below 80% all correspond to experimental situations which are unlikely. The smallest improvements occur when $|\gamma_2| < .1$; since the reparameterization of Chapter III fixed the standard deviation σ of observational error at unity, values of $|\gamma_2| < .1$ should be interpreted as referring to experiments for which $|\gamma_2| < .1 \sigma$. In practice the experimenter will be able to increase the value of $|\gamma_2|$ relative to σ simply by increasing the step size between single observations. As $|\gamma_2|/\sigma$ increases with t^* fixed, the loss incurred by any of the stopping rules decreases, because the expected response function $\eta(t) = \gamma_0 + \gamma_2 t + \gamma_2 t^2$ has a slope which is steeper for small t and which changes more rapidly as $t \rightarrow t^*$, so that the probability of incorrectly diagnosing the slope from the value of an unbiased estimator is lower. Therefore, if the experimenter chooses a large enough step size, he should be able to obtain a 90% improvement in loss by using Rule 2 or Rule 5 at the 25% level, provided he knows the value of σ^2 .

Table VI: Percent Improvement, Rules 2 and 5

t*	γ_1	γ_2	Stand- ard	Mini- mum	Rule 2	%	Rule 5	%
4.0	.40	-.05	7.52	5.25	5.66	82	6.17	59
	.80	-.10	6.29	4.25	4.64	81	4.28	99
	1.20	-.15	5.26	3.04	3.51	79	3.06	99
	2.00	-.25	3.37	1.55	2.39	54	1.75	89
4.7	1.40	-.15	7.17	3.29	3.78	87	3.30	99
5.0	.20	-.02	14.48	9.79	10.21	91	10.41	87
	1.00	-.10	14.46	5.05	5.74	87	5.05	100
5.3	1.60	-.15	9.14	3.33	3.99	89	3.35	99
6.0	.60	-.05	20.06	9.81	10.68	92	9.95	99
	1.20	-.10	15.49	5.47	6.53	89	5.59	99
	1.80	-.15	11.27	3.40	4.17	90	3.58	98
6.7	.40	-.03	28.06	13.79	16.35	82	14.22	97
	2.00	-.15	13.41	3.38	4.21	92	3.87	95
7.0	1.40	-.10	21.09	6.20	7.36	92	6.23	92
8.0	.80	-.05	38.47	12.09	15.52	87	13.37	95
	1.60	-.10	27.09	6.24	8.12	91	6.24	100
9.0	1.80	-.10	32.80	6.52	8.32	93	6.80	99
10.0	.20	-.01	76.31	13.40	51.99	39	44.34	51
	.40	-.02	73.48	16.62	43.45	53	32.99	71
	.60	-.03	70.49	16.57	35.78	64	25.29	84
	1.00	-.05	61.75	15.01	21.99	85	15.51	99
	2.00	-.10	38.58	6.87	8.59	95	7.65	98

The third and most speculative recommendation covers the case in which the experimenter has no previous reliable estimate of the error variance σ^2 . It would seem reasonable to estimate σ^2 from the observations of the first-order experiment and use the estimate s^2 to replace the normally distributed test statistics z_n by z_n/\sqrt{s} , which has a student's t-distribution. Because the resulting t-tests will be less powerful than their normal analogs, the experimenter must take steps to combat the inevitable increase in the losses incurred. One possibility would be to take replicated observations at the center of the first-order design to increase the degrees of freedom associated with the estimate of σ^2 ; another would be to fit a quadratic polynomial in t to the single observations and pool any lack of fit with the estimate obtained from the first-order design. A further difficulty arises in determining the best level for the tests, which will be very highly correlated. Purely for lack of any obvious alternative, the previous level $\alpha=.25$ is suggested.

Recommendation 3: If the experimenter does not have a reliable estimate for σ^2 , he should estimate it from a first-order design augmented at the design center to provide additional degrees of freedom, and base his stopping rule on the following sequence of tests conducted at the 25% level:

$$n=1 \quad \text{Reject } H_{01} \Leftrightarrow y_1 - y_0 < -t_{.25, r} \sqrt{2 \cdot \text{MSE}}$$

$$n \geq 2 \quad \text{Reject } H_{0n} \Leftrightarrow (y_n - y_{n-2})/2 < -t_{.25, r} \sqrt{\frac{\text{MSE}}{2}}$$

IV.C DISCUSSION

Though the search for a better stopping rule can only

be judged a success, the judgment remains subject to qualification on two accounts: the new rule is still not optimal in any rigorous sense, and it represents an improvement of but one phase of a larger procedure which is itself open to question. Regarding the first qualification, one cannot escape the feeling that further exploration would yield further improvements. Possibly α should not be held at a fixed level throughout the sequential decision phase, but varied from a relatively high value, say .60, for the first test, to a value below .25, possibly as low as .10, as the number of single observations increases beyond six or seven. Possibly the best level for use with t-tests differs from the proposed $\alpha=.25$; certainly one could gather data to support or deny this possibility. Finally, there is no reason to constrain the experimenter to equally spaced observations. It should be possible to obtain early estimates of t^* and space the single observations accordingly.

Even if all the preceding possibilities were explored and exploited, one would still have to contend with the second qualification, which questions the value of applying steepest ascent to statistical problems. Following the strategy outlined for the statistician's problem in Chapter I, this thesis has developed and applied an optimality criterion to a small segment of the experimenter's problem. To whatever extent possible, it should conclude by regrafting the excised subproblem onto the parent problem, and reintegrating the limited

optimality criterion into the larger context of the optimal solution to that parent problem. In particular, it should consider the effect of the new stopping rule on the number of observations required by the response surface experimenter to determine the region of optimum response, and attempt to compare the new variant of steepest ascent with its alternative, the simplex method. For lack of data, any evaluation of the new variant must necessarily be speculative, but one can at least construct examples which illustrate the strengths and weaknesses of the new variant relative to those of the old variant and the simplex method. The examples thus suggest a context for a more rigorous comparison of region-seeking methods.

The distinction between region-seeking methods and maximum-seeking methods is crucial to any evaluation, and should be reiterated, lest the reader confuse the proposed evaluation with a previous study by Brooks ([8]). The region-seeking method is a preliminary phase of those response surface experiments in which the first design region is distant from the optimum. The maximum-seeking method is always the final phase of the experiment; in those experiments which require region-seeking, it completes the job which the region-seeking method begins. In practice, particularly in the past, the distinction has been not so much between techniques as between purposes: many of the techniques, including the

simplex method and the older variant of steepest ascent, can be used for both region-seeking and maximum-seeking, so that the experimenter does not always distinguish the two.

The Brooks article evaluates only maximum-seeking methods. Each of several methods is applied to the problem of maximizing the response associated with each of four surfaces judged to be typical (although all four surfaces have absolute maxima--there are no saddles or rising ridges). The efficiency of each method is measured by the magnitude of the expected response at that point that the method determines to be optimum. Brooks reports that the variant of steepest ascent mentioned in the footnotes of Chapter I performs better than the other methods considered; Spendey, et. al. ([22]) apply his criterion to the simplex method and find it better than the steepest ascent variant.

The Brooks criterion for measuring the efficiency of maximum-seeking procedures cannot be applied to region-seeking procedures for two reasons. First, none of the four response surfaces used covers a large enough region of the factor space to allow region-seeking. Each experiment of necessity originates in a region near the optimum. Second, the measure of efficiency is inappropriate because it ignores the number of observations required by a method to move close to the optimum. A hypothetical first-order method that needed only ten observations to obtain a response that was 90% of the maximum might nevertheless receive a low efficiency rating if heavy biasing prevented it from obtaining a response above 95%

of the maximum after another twenty observations. Despite the method's low rating for maximum-seeking, it might be efficient for region-seeking.

Any measure of region-seeking efficiency must incorporate the number of observations required by a method to find the target region it seeks. If the target region were defined to be any part of the factor space where the expected response exceeded some large fraction (e.g., 90%) of the maximum, then any quantity inversely proportional to the number of observations needed to reach the target region could be used to measure efficiency. Alternatively, a cost to distance ratio could be formed by dividing the number of required observations by the component of the distance travelled in the direction of the optimum. Small values of the cost to distance ratio would indicate high efficiency. The latter criterion has the advantage of depending less than the former on the nature of a particular response surface because it does not require an arbitrary definition of the target region.

The following example uses the cost-to-distance ratio to illustrate the advantage of using the proposed new variant of steepest ascent based on Rule 2 instead of the old variant. Suppose that in a two-dimensional experiment the path of steepest ascent is estimated from the four observations of a 2^2 factorial design. If the expected response for points on the path is determined by the parameters $\gamma_1 = 2.0$, $\gamma_2 = -.1$ (relative to σ), an experimenter using the new variant of

steepest ascent can expect to take 11 single observations (to the nearest integer), stopping at $t=10$; but an experimenter using the old variant can expect to take only 5 observations, stopping at $t=4.5$. The respective measures of efficiency are

$$\text{New variant: } \frac{(4+11) \text{ obs.}}{10 \text{ units}} = 1.5 \text{ obs. per unit,}$$

$$\text{Old variant: } \frac{(4+5) \text{ obs.}}{4.5 \text{ units}} = 2.0 \text{ obs. per unit.}$$

If the number of factors were greater, so that the cost of the first-order design increased relative to the cost of the single observations, the saving to an experimenter using the new variant would increase accordingly. For example, in a four-factor experiment, the ratios for the two variants would become

$$\text{New variant: } \frac{(16+11) \text{ obs.}}{10 \text{ units}} = 2.7 \text{ obs. per unit,}$$

$$\text{Old variant: } \frac{(16+5) \text{ obs.}}{4.5 \text{ units}} = 4.67 \text{ obs. per unit.}$$

It is not possible to include the simplex method in the comparison using the data now available without abandoning the quantitative measure of efficiency, but a qualitative comparison can be made. Because the simplex method is able to follow a non-linear path of ascent (Figure 6 of Chapter I), the new variant of steepest ascent would probably be less efficient than the simplex for situations which, like Example I.1, allow the experimenter to begin less than ten units from the

optimum in a region where the curvature is already appreciable. But because the method of steepest ascent adheres more closely to a linear path of ascent, and allows the experimenter more freedom in the location of the points at which he can take single observations, the new variant of steepest ascent would probably be more efficient for plane-climbing. Thus the optimum region-seeking procedure might combine the two methods, beginning with steepest ascent and converting to the simplex method as soon as an appropriate measure of curvature reached some critical value.

The questions left unanswered by the necessarily imprecise comparison of the region-seeking procedures emphasize the need for further investigation. It should be possible, using a computer, to determine which response surfaces are best explored by the simplex method, which by steepest ascent. If the results indicate that steepest ascent is more efficient than the simplex method for some situations, it should also be possible to determine criteria which allow the experimenter to recognize those situations in order to make the most efficient use of both methods.

REFERENCES

- [1] Arnold, Harvey J., Bucher, Bradley D., Trotter, Hale F., and Tukey, John W., "Monte Carlo techniques in a complex problem about normal samples," in Meyer, Herbert A., ed., Symposium on Monte Carlo Methods. New York: John Wiley and Sons, 1956:80-88.
- [2] Box, G.E.P., "The exploration and exploitation of response surfaces: some general considerations and examples," Biometrics 10 (1954): 16-60.
- [3] ---, and Coutie, M.A., "Application of digital computers in the exploration of functional relationships," Proc. Inst. Elec. Engrs. 103 (1956): Part B, Supplement No. 1: 100-107.
- [4] ---, and Draper, Norman R., "A basis for the selection of a response surface design," J. Amer. Stat. Assoc. 54 (1959): 622-654.
- [5] ---, and Draper, Norman R., "The choice of a second-order rotatable design," Biometrika 41 (1954): 190-199.
- [6] ---, and Hunter, J.S., "Multifactor designs," Annals of Math. Stat. 28 (1957): 195-241.
- [7] ---, and Wilson, K.B., "On the experimental attainment of optimum conditions," J. Roy. Stat. Soc., Series B, 13 (1954): 1-45.
- [8] Brooks, S.H., "A comparison of maximum seeking methods," Operations Research 7 (1959): 430-457.
- [9] ---, and Mickey, M.R., "Optimum estimation of gradient direction in steepest ascent experiments," Biometrics 17 (1961): 48-56.
- [10] Davies, O.L., The Design and Analysis of Industrial Experiments. New York: Hafner, 1954, Chapter 11.
- [11] Draper, Norman R., and Lawrence, Willard E., "Designs which minimize model inadequacies: cuboidal regions of interest," Biometrika 52 (1965): 111-118.

- [12] Gardiner, D.A., and Cowser, K.E., "Optimization of radionuclide removal from low-level process wastes by the use of response surface methods," Health Physics 5 (1961): 70-78.
- [13] Gaylor, D.W., and Merrill, J.A., "Augmenting existing data in multiple regression," Technometrics 10 (1968): 73 ff.
- [14] Graybill, Franklin A., An Introduction to Linear Statistical Models, vol. 1. New York: McGraw Hill, 1961.
- [15] Hader, R.J., Karson, M.J., and Manson, A.R., "Minimum bias estimation and experimental design for response surfaces," Technometrics 11 (1969): 461-476.
- [16] Kahn, Herman, "Use of different Monte Carlo sampling techniques," in Meyer, Herbert A., ed., Symposium on Monte Carlo Methods. New York: John Wiley and Sons, 1956: 146-190.
- [17] ---, "Multiple quadrature by Monte Carlo methods," in Ralston, Anthony, and Wilf, Herbert S., eds., Mathematical Methods for Digital Computers, vol. 1. New York: John Wiley and Sons, 1960: 249-257.
- [18] Karson, M.J., "Design criterion for minimum bias estimation of response surfaces," J. Amer. Stat. Assoc. 65 (1970): 1565-1572.
- [19] Myers, Raymond H., Response Surface Methodology. New York: Allyn and Bacon, 1971.
- [20] Ralston, Anthony, A First Course in Numerical Analysis. New York: McGraw Hill, 1965: 441-442.
- [21] ---, "Methods for numerical quadrature," in Ralston, Anthony and Wilf, Herbert S., eds., Mathematical Methods for Digital Computers, vol. 1. New York: John Wiley and Sons, 1960: 242-248.
- [22] Spendley, W., Hext, G.R., and Himsworth, F.R., "Sequential application of simplex designs in optimisation and evolutionary operation," Technometrics 4 (1962): 441-461.

- [23] Stroud, A.H., "Approximate multiple integration," in Ralston, Anthony, and Wilf, Herbert S., eds., Mathematical Methods for Digital Computers, vol. 2. New York: John Wiley and Sons, 1967: 145-155.
- [24] Trotter, Hale F., and Tukey, John W., "Conditional Monte Carlo for normal samples," in Meyer, Herbert A., ed., Symposium on Monte Carlo Methods. New York: John Wiley and Sons, 1956: 64-79.
- [25] Webster, J.T., "On the method of Das in evaluating a multivariate normal integral," Biometrika 57 (1970): 657-664.
- [26] Wilf, Herbert S., "Advances in numerical quadrature," in Ralston, Anthony and Wilf, Herbert S., eds., Mathematical Methods for Digital Computers, vol. 2. New York: John Wiley and Sons, 1967: 133-144.

APPENDIX I

CRUDE DISTRIBUTIONAL PROPERTIES OF THE
WEIGHTED LEAST SQUARES ESTIMATES OF II.C

Computer simulation provided rough approximations to the mean and variance of the estimated angle $\hat{\phi} = \tan^{-1}(\hat{\beta}_2/\hat{\beta}_1)$ of the line of steepest ascent for various values of relative to the smallest observation y_{min} . Normal random numbers with controlled variance were appended to the expected response values of the two examples and the weighted estimates computed in each of 1000 trials. Figures 26-29 plot the mean and variance of $\hat{\phi}$ as a function of α , for four different values of σ , the standard deviation of the observations. The mean angle $\hat{\phi}$ appeared to depend only slightly on σ^2 , but considerably, of course, on α . The variance, as expected, was inordinately high in the vicinity of the singularities of the rational function; high variance effectively eliminates values of α much below $-y_{min}$ as possible choices for weights.

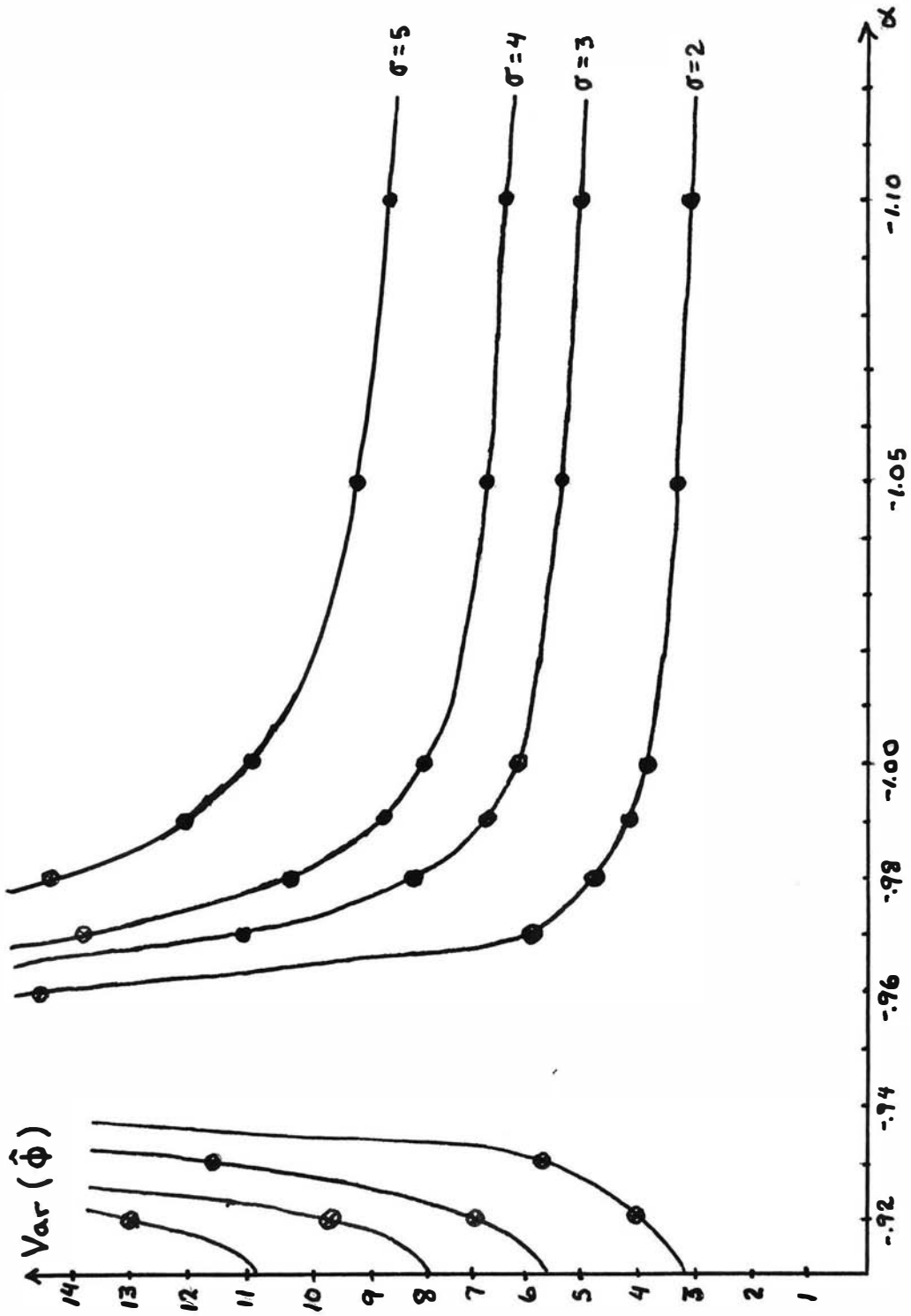


Fig. 28 Variance of $\hat{\phi}$ as a function of α ; Ex. II. 2

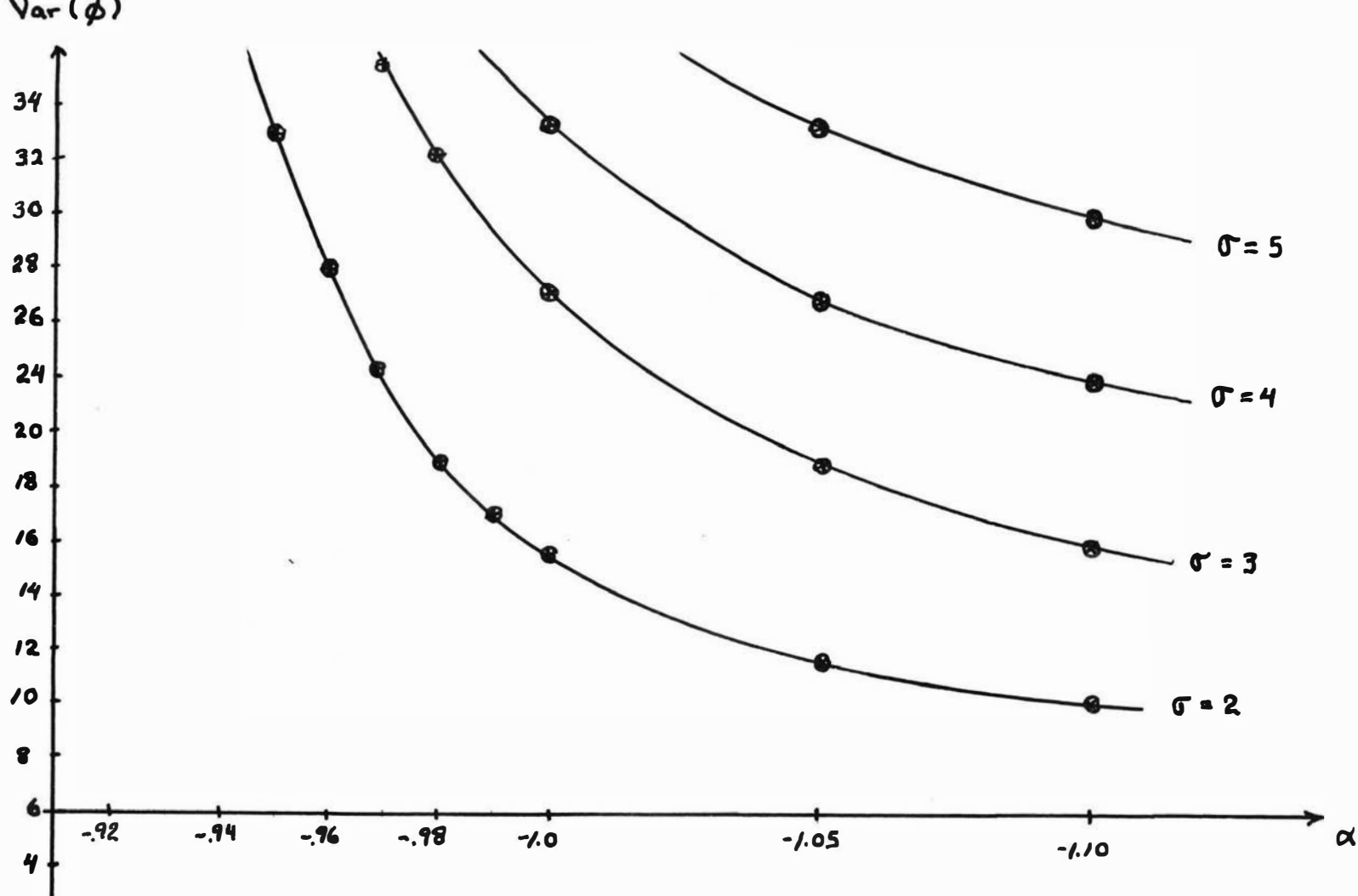


Fig. 27 Variance of $\hat{\phi}$ as a function of α : Ex. II.3

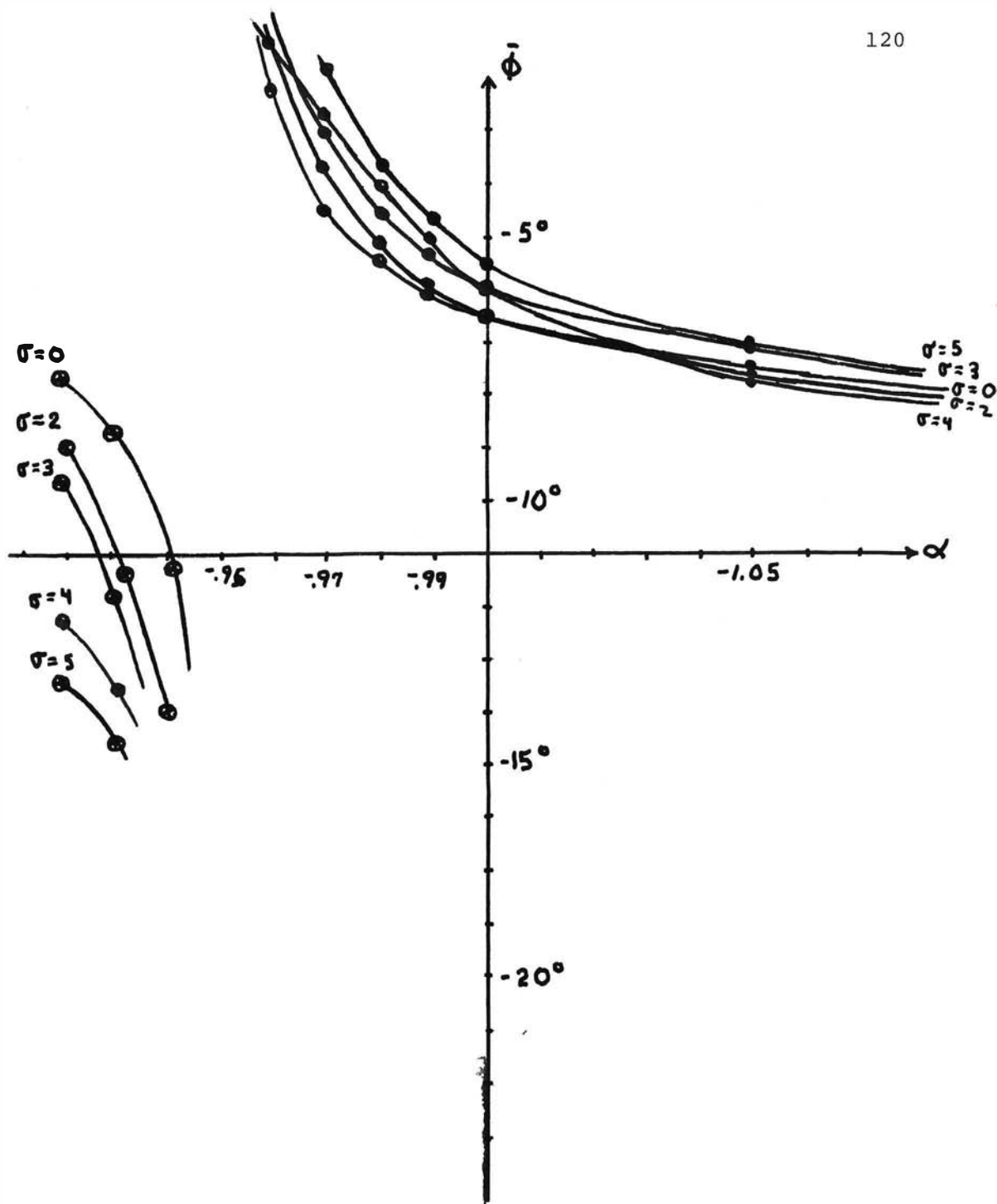


Fig. 28 Mean of $\hat{\phi}$ as a function of α : Ex. II. 2

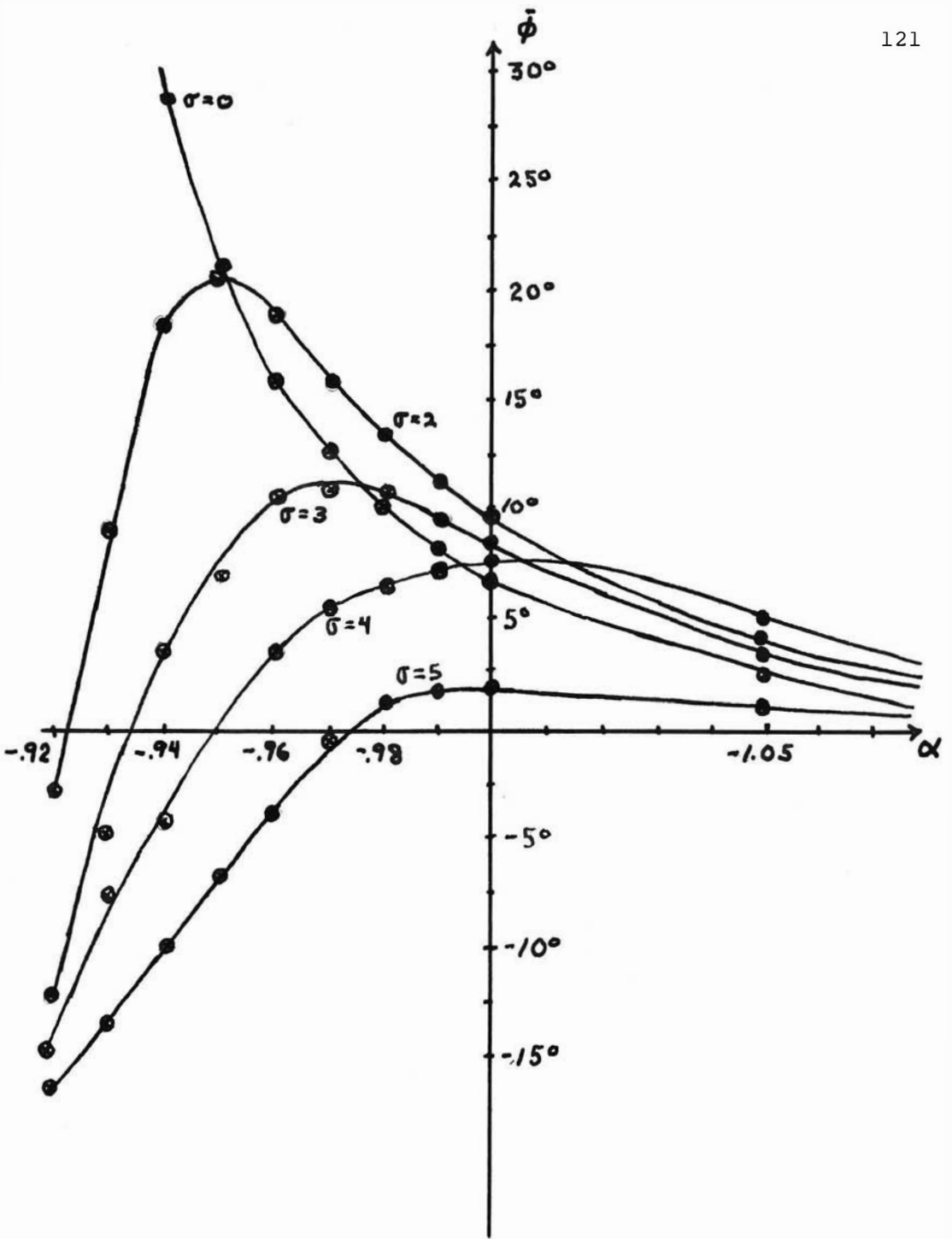


Fig. 29 Mean of $\hat{\phi}$ as a function of α : Ex. II. 3

APPENDIX II: THE MULTIVARIATE INTEGRAL PROBLEM

The multivariate integral problem arises in this thesis from two diverse statistical contexts and calls upon equally diverse numerical methods for its solution. In its most general form the problem requires evaluation of probabilities associated with particular semi-infinite rectangular regions under a multivariate normal density:

Given $\underline{z} \sim N_n(\underline{\xi}, V)$, find

$$P(a, n | \underline{\xi}, V) = \Pr \{ z_i > a : i=1, 2, \dots, n \}$$

$$= \int_a^\infty \cdots \int_a^\infty f(\underline{z} | \underline{\xi}, V) d\underline{z},$$

$$\text{where } f(\underline{z} | \underline{\xi}, V) = \frac{|V|^{-1/2}}{(2\pi)^{n/2}} \exp\left\{-\frac{1}{2}(\underline{z} - \underline{\xi})' V^{-1}(\underline{z} - \underline{\xi})\right\}.$$

Two broad classes of numerical methods can be applied to the computation of such integrals: numerical quadrature and Monte Carlo. The preferred method depends on the particular context surrounding the numerical problem, so that the method of evaluation chosen here does not claim to be appropriate

under all circumstances, but only within the limited context of the second statistical problem described below.

First context. The evaluation of weighted regression as modified in II.C requires computation of the probabilities associated with all possible permutations of a sample of independent normal observations with unit variances and unequal means:

$$\text{Given } y \sim N_n(\underline{\mu}, I), \text{ find } \Pr\{y_1 > y_2 > \dots > y_n\}.$$

By transforming the independent observations to obtain dependent statistics formed from differences, the problem can be restated in terms of the $P(a, n | \underline{\xi}, V)$.

Let $z_i = y_{i+1} - y_i$; then $\underline{z} \sim N_{n-1}(\underline{\xi}, V)$, where

$$\underline{\xi} = \begin{bmatrix} \mu_2 - \mu_1 \\ \mu_3 - \mu_2 \\ \vdots \\ \mu_n - \mu_{n-1} \end{bmatrix}, \quad V = \begin{bmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & 0 & \dots & 2 \end{bmatrix}.$$

$$\begin{aligned} \text{So } \Pr\{y_1 > y_2 > \dots > y_n\} &= \Pr\{z_i > 0 : i=1, 2, \dots, n-1\} \\ &= P(0, n-1 | \underline{\xi}, V). \end{aligned}$$

Second context. The random variable s (stopping distance) of Chapter III requires for its definition the computation of probabilities

$$\begin{aligned} & \Pr \{ \text{Accept } H_{01}, \text{Accept } H_{02}, \dots, \text{Accept } H_{0,n-1}, \text{Reject } H_0 \\ &= \Pr \{ z_1 > z_\alpha, z_2 > z_\alpha, \dots, z_{n-1} > z_\alpha, z_n \leq z_\alpha \}, \\ & \text{where } \underline{z} \sim N_n(\underline{\xi}, V), \quad \text{with } \underline{\xi} \text{ and } V \end{aligned}$$

determined by the particular model and test statistics under consideration. Once again the probabilities can be obtained from the $P(\alpha, n | \underline{\xi}, V)$:

$$\begin{aligned} & \Pr \{ z_1 > z_\alpha, z_2 > z_\alpha, \dots, z_{n-1} > z_\alpha, z_n \leq z_\alpha \} \\ &= \Pr \{ z_i > z_\alpha : i=1, 2, \dots, n-1 \} - \Pr \{ z_i > z_\alpha : i=1, 2, \dots, n \} \\ &= P(z_{\alpha, n-1} | \underline{\xi}^*, V^*) - P(z_{\alpha, n} | \underline{\xi}, V), \quad \text{where} \end{aligned}$$

$\underline{\xi}^*$ and V^* are obtained from $\underline{\xi}$ and V by truncation at $n-1$.

Evaluation of the $P(\alpha, n | \underline{\xi}, V)$ by quadrature methods is feasible only if n is small. Each integral is approximated by an n -fold summation known as a Cartesian product formula ([23]):

$$P(\alpha, n | \underline{\xi}, V) = \int_a^\infty \int_a^\infty \cdots \int_a^\infty f(\underline{z} | \underline{\xi}, V) d\underline{z}$$

$$\approx \sum_{i=1}^m \sum_{j=1}^m \cdots \sum_{t=1}^m B_i B_j \cdots B_t f(\underline{w}_\tau | \underline{\xi}, V)$$

where the \underline{w}_τ 's ($\tau=1, 2, \dots, m^n$) are selected points in the space of the random variables \underline{z} and the B 's are appropriate weights. For example, one might use the B 's and \underline{w} 's designated by a Gauss-Legendre formula which is exact for integrals of polynomials of degree $\leq 2m-1$ over a finite hypercube.

Unfortunately, the number of terms in the summation formula increases geometrically in n , the dimension of the multiple integral, so that even when the number m of grid points is small, the quadrature formula is efficient only for small n . For example, if $m=5$ and $n=8$, the n -fold summation contains $5^8 = 3.9 \times 10^5$ terms.

The dimension of the integral can sometimes be reduced by as much as a factor of two using the method of

Das, provided the minor determinants of V satisfy three sets of conditions ([25]). Under these conditions, V may be written as $c^2 I + BB'$, where B is $n \times k$, $k \leq (n+1)/2$ and the n -variate integral may be rewritten as a k -variate integral by performing a suitable linear transformation on V . None of the dispersion matrices V arising from the two statistical contexts of this thesis satisfy the requisite conditions for the method of Das, and no reduction of dimension is possible. Because some of the integrals involve as many as sixteen dimensions, the failure to achieve a reduction in dimension rules out the use of quadrature methods.

The alternative to quadrature methods for evaluating multivariate integrals is based on random sampling. In its crudest form the method approximates each $P(a, n | \xi, V)$ by generating random samples from the density $f(\underline{z} | \xi, V)$ and forming the ratio of successes to trials, where a success is any observation \underline{z} for which $\{z_i > a : i = 1, 2, \dots, n\}$. The law of large numbers insures that the ratio will converge to the probability it estimates as the sample size $N \rightarrow \infty$, and the theory for binomial distributions gives the variance of the estimator as

$$\text{Var}(\hat{p}) = \frac{1}{N} \{ P(a, n | \xi, V) [1 - P(a, n | \xi, V)] \}$$

Numerous refinements are available ([16]) to accelerate the rate at which the ratio converges to the probability it estimates; when any of these refinements are used to modify the straightforward random sampling the resulting method of approximation is known as Monte Carlo.

If one were evaluating a small number of unrelated probabilities $P(\alpha, n | \xi, V)$ the refinements known generically as Russian Roulette and Splitting could be applied to the problem with a resulting increase in efficiency, but at the cost of distorting the sampling procedure so that each probability would have to be estimated from a separate sample. For the integrals whose values determined the probabilities associated with the stopping distance s , it is more efficient to use a straightforward sampling scheme which, though less efficient for estimating single probabilities, allows simultaneous estimation of a large class of probabilities.

In particular, from a single random vector $\xi' = (\xi_0, \xi_1, \dots, \xi_n)$, $\xi \sim N_{n+1}(\theta, I)$ one can generate a vector of correlated errors $\zeta' = (\xi_1 - \xi_0, \dots, \xi_n - \xi_{n-1})$ to be added to the expectations of the test statistics $\underline{z}' = (y_1 - y_0, y_2 - y_1, \dots, y_n - y_{n-1})$. Although the expectations $E \underline{z}$ depend on the model parameters γ_1 and γ_2 , the same random errors can be added to each vector of expectations so that probabilities for all pairs (γ_1, γ_2) may be computed at once. Further, one can estimate $P(\underline{z}_\alpha, n | \xi, V)$ simultaneously for several values of \underline{z}_α .

Theoretically, it would have been possible to estimate from a single sample all of the 48,400 probabilities needed to evaluate the five stopping rules at five test levels for 121 parameter pairs, but the limited storage capabilities of the IBM 1130 computer required the use of five separate samples. A sample size of $N=10,000$ guaranteed that all estimates had variances less than 6.25×10^{-6} .